ESTUDIO DE LA DEGRADACIÓN DEL CAMPO SOLAR

STUDY OF SOLAR FIELD DEGRADATION

TRABAJO FIN DE MÁSTER
CURSO 2020-2021

AUTOR
LUIS JULIÁN GARCÍA GIMÉNEZ

DIRECTOR
RAFAEL CABALLERO ROLDÁN

COLABORADOR
MANUEL LADRÓN DE CEGAMA

MÁSTER EN INTERNET DE LAS COSAS
FACULTAD DE INFORMÁTICA
UNIVERSIDAD COMPLUTENSE DE MADRID

ESTUDIO DE LA DEGRADACIÓN DEL CAMPO SOLAR

STUDY OF SOLAR FIELD DEGRADATION

TRABAJO DE FIN DE MÁSTER EN INTERNET DE LAS COSAS
DEPARTAMENTO DE SISTEMAS INFORMÁTICOS Y COMPUTACIÓN

AUTOR
LUIS JULIÁN GARCÍA GIMÉNEZ

DIRECTOR
RAFAEL CABALLERO ROLDÁN

COLABORADOR
MANUEL LADRÓN DE CEGAMA

CONVOCATORIA: SEPTIEMBRE 2021
CALIFICACIÓN: 7,5

MÁSTER EN INTERNET DE LAS COSAS
FACULTAD DE INFORMÁTICA
UNIVERSIDAD COMPLUTENSE DE MADRID

6 DE SEPTIEMBRE DE 2021

DEDICATORIA

A mi familia por creer siempre en mí.

AGRADECIMIENTOS

A mi familia por apoyarme siempre: a mis padres, a mis hermanos y a mi abuela.

Gracias por darme fuerzas en los momentos más duros, sin vosotros no habría sido posible.

Muchas gracias, a mi tutor Rafa por su ayuda y consejos en el desarrollo del trabajo.

Y en general a todos los profesores de este Máster, por transmitirme sus conocimientos.

Muchas gracias Sara por el apoyo constante.

A todos, muchas gracias.

RESUMEN

La energía no renovable se ha utilizado durante muchos años, pero los recursos de la Tierra son limitados. La energía no renovable provoca daños sobre el planeta favoreciendo al cambio climático, por lo que en las últimas décadas se está apostando por las energías renovables.

La energía renovable es la solución ante muchos problemas que se causan en el medio ambiente utilizando energía no renovable. Una de las energías renovables más punteras es la energía fotovoltaica la cual permite obtener energía eléctrica a través de la radiación solar.

A pesar de que la energía solar es una energía limpia que no produce contaminación, la vida útil de los paneles solares es limitada ya que estos se van degradando con el paso del tiempo.

Por tanto, es interesante conocer el rendimiento de los campos solares para conocer cuánto se han degrado los campos solares y saber si están cumpliendo la degradación anual establecida en la garantía del fabricante. Esto va a ayudar a no desechar módulos que aún funcionen correctamente, y así evitar que se creen gran cantidad de residuos.

Como la energía solar depende en gran medida de factores ambientales, se utilizan sensores para conocer datos de radiación y temperatura. Estos sensores van a permitir predecir cuál es la potencia antes de instalar una planta solar. En este trabajo se muestra como, utilizando modelos de aprendizaje automático se puede obtener el rendimiento de los campos solares utilizando gran cantidad de datos muestreados con los sensores.

Además se va a presentar un método que permite mediante el análisis de datos, comparar el rendimiento entre varios años y estudiar si esa diferencia corresponde con la degradación del campo solar.

Palabras clave

Energía renovable, degradación, energía solar, fotovoltaico, machine learning, predicción, tratamiento de datos, big data.

ABSTRACT

Non-renewable energy has been used for many years, but the Earth's resources are limited. Non-renewable energy causes damage to the planet and contributes to climate change. Therefore, in recent decades, renewable energy has been the solution to many problems.

Renewable energy is the solution to many problems that are caused in the environment using non-renewable energy. One of the most advanced renewable energies is photovoltaic energy, which allows to obtain electrical energy through solar radiation.

Although solar energy is a clean energy that does not produce pollution, the useful life of solar fields is limited because they degrade over time.

Thus, it is interesting to know the performance of the solar fields to know how much the solar fields have degraded and to know if they are complying with the guarantee indicated by the manufacturer. This will help to avoid discarding modules that are still working properly, and thus avoid creating a large amount of waste.

As solar energy depends to a large extent on environmental factors, sensors are used to determine radiation and temperature data. These sensors will make it possible to predict the power output before installing a solar plant. By means of machine learning models, the performance of solar fields can be obtained using a large amount of data sampled with the sensors.

On the other hand, by processing and analyzing the data, we compare the performance between several years and study whether this difference corresponds to the degradation of the solar field.

Keywords

Renewable energy, power, degradation, solar energy, photovoltaic, machine learning, prediction, data processing, big data

ÍNDICE DE CONTENIDOS

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

Capítulo 1 - Introducción

En este proyecto se va a explicar cómo funciona la energía solar a grandes rangos, dando a conocer conceptos teóricos que son necesarios para el estudio de la degradación del campo solar. Además, se va a explicar cómo se ha implementado el estudio práctico en Python, explicando más en detalle el proceso que se ha seguido en cada fase. Este experimento va a consistir en utilizar modelos para predecir el rendimiento de los campos solares y utilizar técnicas de tratamiento de datos masivas para obtener resultados de la magnitud con la que se degrada el campo solar.

1.1 Motivación

En los últimos años, las energías renovables están continuamente creciendo, y se prevé que esto siga siendo así en los próximos años. Ya que el impacto que provocan las energías no renovables sobre el medio ambiente está provocando que estas se utilizan cada vez menos.

Dentro de las energías renovables, una de las más destacadas es la energía solar, la cual está aumentando notablemente su participación en la generación de energía eléctrica.

El material necesario para crear energía solar cada vez es más accesible, incentivando a las personas a instalar módulos solares en sus viviendas. Además, en los últimos años está disminuyendo el precio de los módulos solares para promover esta energía.

Como los módulos se van degradando con el paso de los años, es conveniente seleccionar módulos de buenos materiales a la hora de realizar la compra, ya que si no puede ser que los módulos se degraden con mayor velocidad.

Los fabricantes de módulos solares deben de cumplir una garantía de rendimiento donde indica que el módulo generará unos valores mínimos de rendimiento.

Con el fin de contaminar lo menos posible, se deben comprar los módulos que mayor vida útil tengan y menos se degraden para generar los mínimos residuos posibles.

Para verificar que la garantía del fabricante se cumple, es necesario realizar un estudio de la degradación del campo solar.

Aunque estos estudios no se suelen realizar en plantas solares domésticas, sí se suelen realizar en grandes empresas que colaboran con la red eléctrica del país. Estas empresas suelen tener grandes plantas solares que generan mucha potencia y les resulta interesante conocer cuál es la degradación del campo solar. Ya que la pérdida de rendimiento en plantas solares grandes, provoca una pérdida económica muy significativa.

Por ello, las empresas generadoras de energía solar estudian la degradación de los módulos comprados para saber si el fabricante cumple la garantía y si no la cumple reclamar para sustituirlos. Así que, estas empresas tienen sistemas de monitorización para muestrear los datos de la planta solar y así poder realizar estudios de la degradación o predicciones del rendimiento de la planta en un futuro. Lo cual resulta muy interesante para estas empresas.

Y en este Trabajo Final de Máster se va a realizar este estudio de la degradación del campo solar y un estudio predictivo de su rendimiento con los datos de una empresa.

1.2 Objetivos

El objetivo de este Trabajo Final de Máster es realizar un estudio de la degradación del campo solar utilizando la gran cantidad de datos que una empresa ha proporcionado. Estos datos incluyen datos adquiridos durante varios años, tanto de rendimiento de la planta solar como de las condiciones ambientales mediante sensores. Para realizar este estudio práctico se va a utilizar la potencia generada en la planta para comprender cómo evoluciona el rendimiento del campo solar con el paso de los años y así evaluar la degradación entre dos años o la degradación entre el primer y el último año. En este estudio va a ser necesario utilizar técnicas de tratamiento y análisis de datos para poder obtener conclusiones de los más de cinco millones de datos que la empresa ha proporcionado para este proyecto. Pero para entender este estudio de la

degradación es necesario presentar unos conocimientos teóricos previos sobre cómo funcionan los sistemas fotovoltaicos.

En definitiva, el Trabajo Final de Máster se puede desglosar en dos objetivos específicos, que se pueden resumir en:

- Predecir la potencia usando sensores de radiación y temperatura usando modelos de aprendizaje automático.
- Estudiar la degradación de los campos solares de una empresa desde 2013 hasta 2020.

1.3 Estructura

Este documento se estructura en 6 capítulos en total . El primero consiste en un preámbulo del presente Trabajo Final de Máster, y a continuación se presenta el capitulo dos, que es una introducción general de las energías y en concreto de la energía solar. Una vez realizada esta introducción, en el capitulo tres se muestran los conceptos teóricos más importantes que son necesarios para analizar la degradación del campo solar.

En el capítulo cuatro se presenta parte del estudio práctico implementado haciendo hincapié en las 3 fases que se han desarrollado: estructura de la planta, preprocesado y visualización de los datos.

Ya que el quinto capítulo va a continuar con la otra parte del estudio práctico que trata sobre el aprendizaje automático. El cual consiste en predecir el rendimiento de una planta solar utilizando modelos de regresión y estimar el porcentaje de degradación anual que han experimentado los paneles solares.

Estos dos capítulos (cuatro y cinco) pertenecen al caso de estudio práctico que se ha implementado por mi parte, y que son los dos capítulos que mayor esfuerzo han requerido en este estudio, ya que se trata del código que se ha tenido que implementar.

Y por último en el sexto capítulo se exponen las conclusiones obtenidas.

Las figuras son de elaboración propia salvo que se indique explícitamente lo contrario.

Se adjunta el código que se ha implementado para este estudio en el siguiente enlace:

Github: **https://github.com/luisjgar/TFM**

Sin embargo, no se puede adjuntar los datos que han sido utilizados debido al acuerdo de confidencialidad firmado con la empresa que ha cedido los datos.

Capítulo 2 - Energías tradicionales y energías renovables

Las fuentes de energía tradicionales se dividen en energía nuclear y combustibles fósiles: carbón, gas, y petróleo. Estas energías se tratan de energías no renovables y por lo tanto tienen la desventaja de ser finitas. Sin embargo, la energía solar es una energía renovable que siempre está disponible en cualquier parte del mundo.

Las energías tradicionales han sido muy utilizadas durante muchos años, pero actualmente cada vez se utilizan menos debido a que las energías renovables cada vez son más baratas y provocan menos contaminación. La sociedad está dándose cuenta de que el sistema energético tradicional no forma parte del futuro ya que cada vez existen menos recursos en la Tierra para generar energía y además los daños causados en esta son considerables.

Así que, el ser humano está mejorando en la forma de elegir cómo se crea la energía. Esto evita el cambio climático al no deteriorar la atmósfera como se hacía con las energías no renovables.

Entonces para crear la energía se propone utilizar energías renovables, que no dañen el medio ambiente. La energía solar asegura proporcionar esta energía de forma sostenible, sin dañar el medio ambiente. Dado que esta energía no emite gases de efecto invernadero, evitando así el calentamiento global. Se trata de la energía más limpia, ya que es la que menos daños causa sobre el medio ambiente.

La energía solar se considera una de las energías renovables más eficientes para ayudar a frenar el cambio climático y actualmente es una de las más utilizadas junto con la energía eólica.

A diferencia de la energía generada por los combustibles fósiles, la energía solar no libera emisiones peligrosas de dióxido de carbono (CO_2). Esta energía sostenible ayuda a frenar el cambio climático y previene daños al medio ambiente.

Según algunos estudios, la superficie terrestre recibe 120.000 terawatios de radiación solar y con esa cantidad se puede satisfacer 20.000 veces más potencia que la que necesita el planeta entero [1].

Obviamente no se puede utilizar toda la superficie de la tierra para obtener energía, pero se puede apreciar que la energía solar es una de las energías más punteras dentro de las energías renovables, ya que hay informes que muestran que la energía solar y eólica representaron el 10% de la generación mundial de electricidad en 2020. Y que casi la mitad de la electricidad de Alemania ha procedido de la eólica y la solar en 2020. Estas estadísticas se pueden observar en la Figura 1.

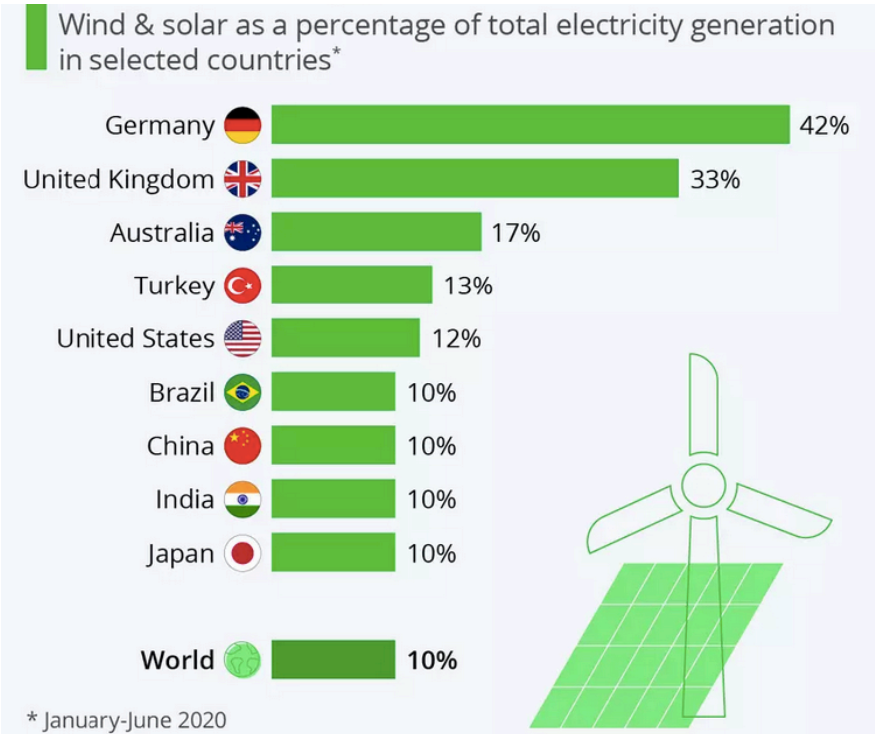

Figura 1. Energía eólica y solar [2]

Aunque en España la energía renovable más utilizada es la eólica, la energía solar está creciendo a un ritmo muy rápido [3]. De hecho, la capacidad solar instalada se ha más que duplicado en los últimos tres años y se han registrado valores máximos históricos en el último año de 2.300 GW/hora [4].

Otra de las ventajas de la energía solar, es que no necesita una gran infraestructura, por lo que existen menos residuos. Pero se debe tener en cuenta que la vida útil de los paneles solares suele ser de unos 25 años. Como esta energía está creciendo en los últimos años, se estima que para 2050 habrán entre 60 y 80 millones de toneladas de residuos de paneles solares en todo el mundo [5].

Para que los paneles solares no generen mucha cantidad de residuos es necesario reciclarlos, ya que su vida es limitada y hay que pensar qué es lo que hay que hacer con ellos cuando se degraden y no sirvan pasados unos 25 o 30 años. ¿Cómo se sabe realmente que ya se han degradado y cuánto? A lo mejor pasados los 25 años un panel sigue funcionando, aunque no se obtenga el mismo rendimiento que cuando se compró, ya que se ha degradado con el tiempo. ¿Pero se debe tirar sin saber cuanto se ha degradado porque han pasado x años? No se debería, ya que si el panel sigue funcionando a un rendimiento aceptable se debe evitar generar residuos. Por ello es recomendable que todas las plantas solares dispongan de un sistema de monitorización que permita obtener conclusiones de la degradación del campo solar.

Además es importante también comprar paneles solares a fabricantes que tengan experiencia en el sector y que proporcionen bastantes años de garantía para evitar generar residuos innecesarios. Para saber que el vendedor cumple con la garantía también es necesario estudiar la degradación de los paneles solares.

Como se ha visto, es muy importante estudiar la degradación del campo solar para evitar generar gran cantidad de residuos, y así que la energía fotovoltaica siga siendo una energía limpia.

Capítulo 3 - Marco teórico

En este capitulo se va a introducir teóricamente cómo funcionan los sistemas fotovoltaicos, los tipos de módulos solares y cómo se fabrican. Para ello se deberá entender de qué materiales está compuesta una célula solar. También se va a exponer qué es el efecto fotovoltaico y cuáles son los factores que determinan el rendimiento de los módulos solares.

Esta introducción teórica va a ser imprescindible para poder comprender en el siguiente capítulo la parte práctica del estudio.

3.1 Sistemas fotovoltaicos

Un sistema fotovoltaico es un conjunto de elementos eléctricos que tiene como objetivo transformar la energía solar en energía eléctrica.

Un sistema fotovoltaico está compuesto por diferente material eléctrico y electrónico. El componente principal es el módulo solar que permite captar la radiación a través de las células solares. Esta radiación incidente es transformada en energía eléctrica, en concreto en corriente continua.

Según la aplicación y la cantidad de energía generada, los sistemas fotovoltaicos se clasifican en dos grupos: sistemas fotovoltaicos aislados o sistemas fotovoltaicos conectados a la red (SFCR). Para comprenderlos mejor se va a explicar las características de cada grupo.

3.1.1 Aislados

Los sistemas fotovoltaicos aislados son también conocidos como autónomos y se caracterizan por usar la energía que van almacenando en las baterías. Se trata de sistemas independientes que no están conectados a la red ya que el transporte eléctrico es costoso (suele ocurrir en zonas industriales o rurales).

El sistema está compuesto por los módulos solares que captan la radiación, un regulador que entrega la energía al sistema, un inversor para transformar la energía y baterías para almacenar la energía.

3.1.2 Conectados a la red

Los sistemas fotovoltaicos conectados a la red son sistemas que generan energía para ser administrada a la red eléctrica. Aunque el objetivo principal de estos sistemas es el autoconsumo. Estos sistemas pueden disponer además de un equipo de almacenamiento. Estos sistemas están compuestos por varios campos solares, los cuales son conectados a la red a través de una unidad de acondicionamiento de energía (PCU) y un regulador MPPT.

Aunque puedan existir sistemas con baterías por si suceden desconexiones en el sistema, lo normal es que los sistemas no tengan baterías ya que están conectados directamente a la red que suministra la electricidad como un servicio.

El sistema dispone de contadores para saber la energía que consume y la que produce. Ya que cuando el sistema produce energía de más, esta energía se suministra a la red eléctrica para recibir un beneficio económico. Y por otro lado, si el sistema solicita menos energía que la energía que se genera con la instalación, el sistema consume la energía de la red.

3.2 Elementos de la planta

Los sistemas fotovoltaicos dependiendo de la aplicación y del diseño pueden tener unos elementos u otros distintos. Sin embargo, las instalaciones solares conectadas a la red suelen ser bastante estándar y disponer de los mismos elementos básicos, que son los que se van a explicar a continuación: los módulos solares, los inversores y los sensores.

Aunque existen muchos tipos de elementos incluidos en las plantas solares. En este capítulo se van a presentar los conceptos generales de estos elementos, los cuales son comunes a todas las plantas solares.

3.2.1 Módulo solar

Es el elemento principal de las plantas solares, ya que es el encargado de captar la radiación solar y transformarla en corriente continua. Los módulos solares están compuestos de varias células solares conectas en serie y paralelo. Los módulos suelen tener 60 o 72 células y son fabricados por empresas como: Jinko Solar, Longi Solar, Trina Solar, Canadian Solar, etc. Los mayores fabricantes de módulos solares a nivel mundial son empresas chinas [6].

La tecnología con la que se fabriquen los módulos va a determinar la potencia de salida de los módulos. Ya que los si se mide la potencia de los módulos policristalinos y la potencia de los módulos monocristalinos en las mismas condiciones estándar de test (1000 $\frac{W}{m^2}$ de radiación y 25ºC de temperatura), se observa que no tienen la misma potencia. Los módulos policristalinos alcanzan menos potencia que los módulos monocristalinos. Por este motivo los módulos policristalinos son más económicos que los monocristalinos. Es por ello, que los módulos más usados son los policristalinos.

La elección del módulo es muy importante en el diseño de un sistema fotovoltaico ya que los módulos solares suponen alrededor de un 20 – 30% de la inversión económica de todo el sistema [7].

3.2.2 Inversor

Se trata del elemento que transforma la corriente continua en corriente alterna. Los paneles solares (conjuntos de módulos solares) transforman la energía solar en corriente continua y esta se quiere transformar a corriente alterna para ser suministrada a la red. Por lo tanto el inversor por una parte está conectado a los paneles solares y por otro lado, está conectado a la red eléctrica. Aunque en realidad entre el inversor y los paneles solares suele haber un regulador el cual también se conecta a las baterías por si se utiliza para autoconsumo. Pero no se explican estos elementos ya que el estudio se trata de un sistema únicamente conectado a red, por lo que no se explica ni el regulador ni las baterías.

Otro matiz, es que el inversor no está conectado directamente a la red eléctrica, sino que entre estos dos existe un contador bidireccional. Pero este elemento tampoco

se explica en detalle porque no es necesario para el estudio de la degradación; ya que para este estudio solo interesa la corriente continua que va de los campos solares al inversor. Lo que ocurre desde el inversor hacia la red no es de interés.

3.2.3 Sensores

Los dos sensores más importantes en una planta solar conectada a la red eléctrica son: sensores de temperatura y sensores de radiación.

Estos sensores son importantes ya que la temperatura y la radiación son dos de los factores más relevantes para determinar la potencia de los módulos fotovoltaicos.

Comenzando por los sensores de temperatura, se debe conocer que se clasifican en dos grupos. Unos sensores de temperatura se encargan de medir la temperatura ambiente (T_a), y otros sensores de temperatura se encargan de medir la temperatura del módulo (T_{cell}). La temperatura que resulta más interesante conocer es la temperatura del módulo que es la que va afectar directamente a la potencia.

El sensor que mide la temperatura que tiene el módulo se debe situar en la propia superficie del módulo. Este sensor se suele ubicar en la parte trasera del módulo solar y habitualmente son sensores PT1000 [8].

En cuanto a los sensores de radiación solar o pirómetros, son radiómetros diseñados para medir la irradiación recibida en una superficie.

Se clasifican en dos grupos según en la ubicación en la que se establezcan. Por una parte, se encuentran los sensores de radiación horizontales, que son aquellos que se encuentran apuntando al cielo y paralelos al suelo, es decir en el plano horizontal. Y por otra parte están los sensores de radiación en el plano del panel que se encuentran con cierta inclinación, exactamente la misma que la del panel.

El sensor en el plano del panel recibe mayor radiación que el sensor horizontal debido a la orientación e inclinación, la cual es estudiada para recibir mayor captación solar.

El sensor que más importancia tiene es el sensor del plano del panel, ya que nos indica la radiación que va a recibir la placa fotovoltaica.

Entonces si el sensor del plano del panel solar es el que interesa para el estudio, ¿para qué sirve el sensor del plano horizontal? El sensor del plano horizontal sirve como sensor de confirmación, para comprobar que la radiación recibida es similar a la del otro sensor y por lo tanto que está funcionando correctamente.

A continuación, en la Figura 2 se muestra la curva de potencia, en la que se puede observar que es mayor la potencia calculada con el sensor del plano del panel (línea verde) que la potencia calculada con el sensor del plano horizontal (línea azul).

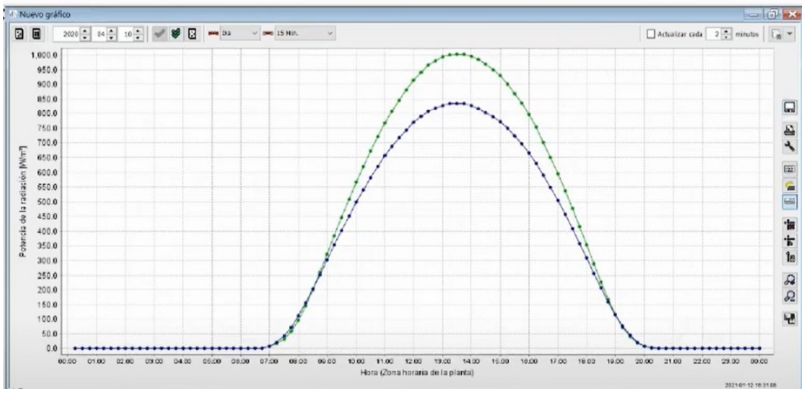

Figura 2. Potencia sensor radiación horizontal vs sensor radiación plano del panel

Como la potencia depende la radiación, la confirma que la radiación es mayor en el plano del panel. La curva de radiación no es la misma si se mide en un sensor horizontal, que si se mide en un sensor en el plano del panel.

3.3 Tecnología de los módulos solares

A día de hoy, la mayor parte de los módulos solares son fabricados en obleas de silicio cristalino [9].

Como se ha mencionado anteriormente, los módulos solares se pueden clasificar según su fabricación en monocristalinos y policristalinos; siendo los módulos monocrostalinos los más eficientes. También existen otras tecnologías como el silicio amorfo o las células de heterounión las cuales están consiguiendo rendimientos

elevados. Pero se van a comentar solo los monocristalinos y policristalinos ya que son las tecnologías más usadas.

En el proceso de fabricación existen unas fases comunes[10]:

La primera fase del proceso de fabricación consiste en obtener el silicio de la mina. Aunque el silicio es un material muy abundante, no se encuentra en estado puro y por lo tanto hay que purificarlo. La siguiente fase consiste en la fabricación de lingotes. En la cual el silicio se funde para obtener obleas planas y posteriormente se crea un lingote que puede ser cilíndrico si se desea obtener silicio monocristalino o cuadrado si se desea obtener silicio policristalino.

En la siguiente fase se cortan los lingotes en láminas finas, también llamadas obleas o *wafers*. Tras realizar procesos químicos sobre estas, se crean las células. Y por último, en la fase final se combinan varias células solares para formar un módulo solar.

Este proceso de fabricación para los módulos monocristalinos y policristalinos se encuentra resumido en la Figura 3.

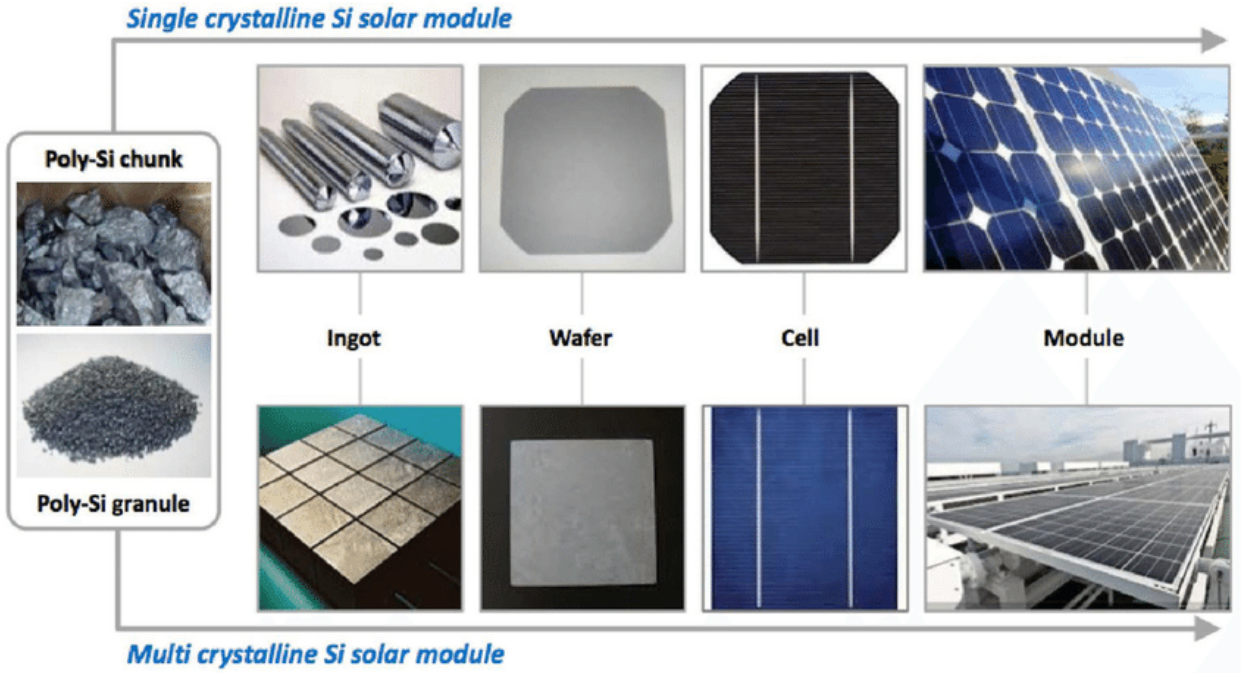

Figura 3. Fabricación módulos monocristalinos (arriba) y policristalinos (abajo) [11]

3.3.1 Monocristalino

Las células de silicio monocristalino son más eficientes que las células policristalinas por su pureza. El rendimiento de las monocristalinas es de alrededor de un 20%. La vida útil de las células monocristalinas es alrededor de 30 años.

La manera más simple de identificar un módulo solar monocristalino de uno policristalino es mirando las esquinas. Ya que las células del monocristalino tienen las esquinas redondeadas y las células policristalinas son rectangulares. Otra forma de diferenciarlas es por el color ya que las monocristalinas tienden a ser más oscuras (casi negras) que las policristalinas (azul marino).

Las obleas monocristalinas se convierten en lingotes mediante el proceso de Czochralski [12]. Este proceso trata de fundir el silicio policristalino a alta presión y temperatura para crear un único cristal monocristalino, el cual es conocido como lingote.

3.3.2 Policristalino

Los módulos de silicio policristalino se fabrican por fundición y moldeado y no se pierde casi material en su fabricación. Aunque los módulos policristalinos sean menos eficientes y tengan una vida útil menor que los monocristalinos, estos son más baratos.

Aunque su eficiencia y su vida útil sea menor, no hay que creer que existe gran diferencia de eficiencia entre ambas, ya que las células policristalinas tienen una eficiencia de alrededor del 15-17%. La vida útil de las células policristalinas es alrededor de 25 años.

Estas células tienen la ventaja que su fabricación es menos costosa energéticamente y se trata de un proceso más rápido.

3.4 Composición célula fotovoltaica

Las obleas de silicio se limpian para eliminar las impurezas tras haberlas cortado. A continuación, las obleas las dopan con fósforo para obtener material de carga negativa (Capa N) y Boro para obtener material de carga positiva (Capa P) y así crear la unión PN [13]. Y seguidamente se limpian los bordes para que no existan cortocircuitos entre las zonas p y n.

A las células solares, se añaden los contactos metálicos utilizando aleaciones de plata y aluminio. El objetivo de los contactos es la transmisión de electrones entre la célula y el cable. Para reducir las pérdidas por reflexión se añade una capa anti reflectante o también llamada anti reflexiva (AR) de óxido de titanio. Y además de esta capa, se puede añadir una textura a la superficie, creando surcos y pirámides para contribuir más a este efecto.

3.5 Efecto fotovoltaico

La célula fotovoltaica es el elemento básico en el que se produce el efecto fotovoltaico. Que se trata de una conversión de la energía solar en energía eléctrica. Es decir una transformación de fotones en electricidad.

El silicio, que es un material semiconductor actúa de aislante a baja temperatura y actúa como conductor cuando recibe mucha energía. Al dopar el silicio se obtienen las capas P y N. Estas 2 capas están separadas por una junta fina para que se mantengan como capas neutras.

Cuando la luz solar incide en el semiconductor, se liberan electrones y se crean huecos. Los electrones son las cargas negativas y los huecos las positivas. El movimiento de estas cargas en direcciones opuestas crea una corriente eléctrica en el silicio que permite que la energía de los fotones de la luz se libere en el semiconductor.

Se evita que las cargas se recombinen, creando así una diferencia de potencial y una corriente eléctrica.

Por lo tanto se puede concluir diciendo que el efecto fotovoltaico se basa en la generación de cargas (electrones y huecos) y la separación de cargas a un conductor que transmita la electricidad.

La corriente de salida de la célula fotovoltaica depende de la eficiencia que esta tenga y de su tamaño, ya que cuanto mayor sea la superficie, mayor número de fotones puede captar, mayor será la cantidad de cargas y mayor será la corriente eléctrica producida.

3.6 Factores determinantes en el rendimiento del módulo solar

Para determinar cuál es la potencia de salida de un campo solar es necesario tener en cuenta muchos factores: la radiación solar, la temperatura del módulo, la orientación e inclinación del módulo, y la degradación del módulo, entre otras cosas [14].

Estos factores afectan al comportamiento de la potencia y por lo tanto se ve reflejado en el rendimiento del módulo solar. Así que es necesario comprender de qué magnitudes depende la potencia y cómo el rendimiento se ve afectado por diferentes factores. Como que el rendimiento varía dependiendo de la posición en la que se establezca el módulo. El rendimiento del módulo también está determinado por la calidad de los materiales de fabricación; ya que si no son de calidad, es más probable que la degradación sea mayor, etc.

3.6.1 Radiación solar

La radiación solar es una energía muy variable y la potencia tiene una relación muy directa con la radiación solar. Por lo que la potencia de salida de un módulo solar no es constante, ya que dependerá de la radiación solar que reciba en ese momento y de otros factores.

También es conocida como irradiación y esta afecta al rendimiento de los módulos solares. Esto se puede ver reflejado en la Figura 4, en la que aparecen las curvas de potencia alcanzando valores distintos dependiendo de la radiación recibida y de la tensión del panel.

Son directamente proporcionales ya que cuando la radiación aumenta la potencia también aumenta, y por lo tanto también aumenta el rendimiento del módulo.

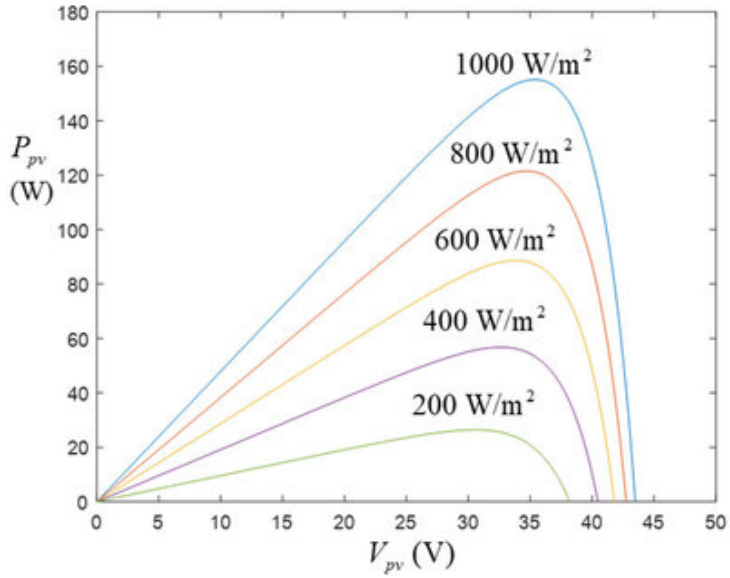

Figura 4. Curvas de potencia que dependen de la tensión y la radiación [15]

Se ha comprobado que la potencia de salida de los módulos depende en gran medida de la radiación solar, por lo que este factor va a ser clave a la hora de estudiar la potencia. Las curvas representadas en la Figura 4 son teóricas, así que se utilizan sensores de radiación que muestrean durante todo el día para estudiar cómo evoluciona la potencia dependiendo de la radiación.

3.6.2 Temperatura del módulo

Las células solares al estar compuestas por silicio, que es un material semiconductor y se ven afectadas por la temperatura.

La potencia es inversamente proporcional a la temperatura. Es decir, en climas cálidos se obtiene menor potencia y en climas fríos mayor potencia.

Por lo tanto, el panel es más eficiente a temperaturas bajas. Ya que a menor temperatura, se obtiene mayor potencia como se observa en la Figura 5.

Esta gráfica muestra las curvas de potencia respecto a la temperatura, cuando se tiene una radiación fija.

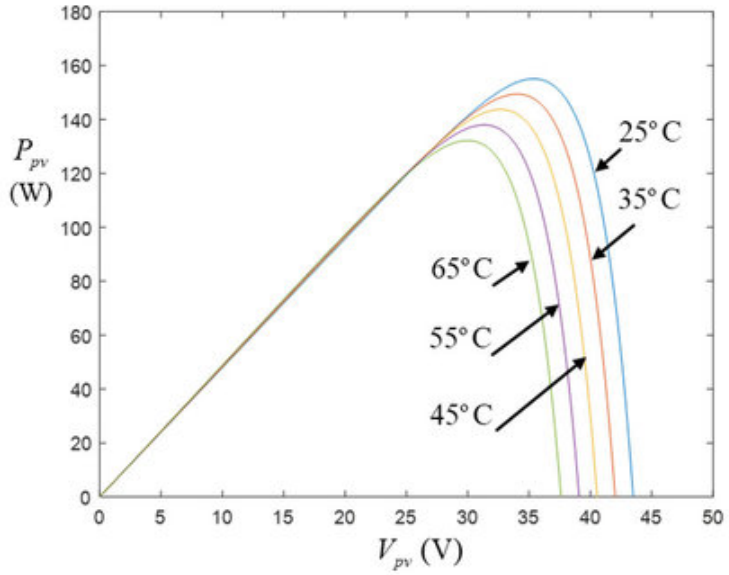

Figura 5. Curvas de potencia dependen de la temperatura [15]

Se puede apreciar claramente en la Figura 5 que para unos valores fijos de tensión, si se observa la línea roja de los 25ºC se obtiene 160 W de potencia y para la línea morada de los 55ºC se obtiene menos potencia: 100 W.

Normalmente las plantas solares se intentan instalar en sitios despejados y fríos. Pero no resulta fácil encontrar ubicaciones con estas características. Estos climas se suelen encontrar en lugares elevados, donde se cumple que el entorno esté despejado y que la temperatura ambiente sea menor.

Aunque en la Figura 5 se muestre la curva de la potencia respecto de la temperatura ambiente, realmente la temperatura que resulta más interesante conocer es la temperatura del módulo que es la que va a afectar directamente a la potencia.

La temperatura del módulo no tiene por qué ser la misma que la temperatura ambiente, ya que existen factores como el viento que afectan a la temperatura.

Por lo tanto, a la hora de buscar una ubicación para la instalación, se buscan sitios elevados, con viento que hace disminuir la temperatura del módulo respecto a la ambiente y con temperaturas bajas; para obtener una mayor eficiencia del módulo solar.

3.6.3 Orientación e inclinación

La orientación ideal de los módulos solares en España es hacia el sur. Esta posición puede obtenerse mediante la hora del día, ya que el sur corresponde con las 12 del mediodía. Por lo que a mediodía es cuando se obtiene el mayor rendimiento [16].

Existen seguidores solares que permiten aumentar el rendimiento de los módulos solares, pero estos suponen un mayor coste para el sistema. Por lo que suelen estar montados sobre estructuras fijas orientadas al sur y con una inclinación específica que maximice la energía generada.

La inclinación se obtiene dependiendo del país en el que se instale el sistema fotovoltaico. Cuanto más cerca esté ese país del ecuador, el ángulo de inclinación de los paneles es menor ya que la radiación es más perpendicular.

En España la inclinación de los módulos varía entre los 20 y 40 grados. El sur de España que se acerca más al ecuador necesita menos ángulo de inclinación (cerca de los 20 grados) y el norte de España que se aleja más del ecuador necesita más ángulo de inclinación (cerca de los 40 grados)

Pero por lo general, en España el grado de inclinación es entre 20 y 35 grados.

También hay que tener en cuenta la orientación y la inclinación para que no existan sombras de algún objeto (edificio, árbol, etc.) sobre la superficie del módulo.

3.6.4 Ensuciamiento

El efecto de ensuciamiento también conocido como "*soiling*" es causado por agentes externos: acumulación de nieve, suciedad, polvo, hojas, polen, excrementos de aves en los paneles fotovoltaicos, etc. [17]. Cuando existe suciedad en la superficie del módulo la potencia se ve afectada negativamente ya que no le llega toda la radiación de forma transparente. Es decir, disminuye la potencia debido al efecto opaco que crea la suciedad o al obstáculo posado sobre la superficie, haciendo así disminuir el rendimiento del módulo fotovoltaico.

La situación empeora aún más cuando no es suciedad, sino nieve que suele cubrir completamente la superficie del módulo; en este caso la potencia es nula y no se produce energía.

Realmente el efecto *soiling* es temporal ya que esta suciedad o elementos sobre la superficie, pueden ser extraídos limpiando los módulos, y una vez realizado el mantenimiento de los módulos, estos vuelven a recuperar su rendimiento habitual.

La Figura 6 muestra la tasa de ensuciamiento durante todos los meses de un año. Cuando esta tasa es igual a uno, significa que no existe *soiling* y que el módulo está funcionando con un buen rendimiento. Y cuando esta tasa disminuye significa que si existe efecto de ensuciamiento por algún agente externo [18].

En la Figura 6 se muestran la tasas de ensuciamiento para diferentes plantas solares, donde cada planta corresponde con una línea de color. Se puede apreciar que todas ellas tienen la misma forma de diente de sierra. Esto es porque hay un cambio brusco de tener ensuciamiento a no tenerlo. Esto es debido al mantenimiento de las plantas solares. Ese cambio brusco aparece cuando los módulos solares son limpiados por el equipo de mantenimiento o por las lluvias y se vuelve a obtener buen rendimiento y la tasa de ensuciamiento vuelve a ser uno de nuevo.

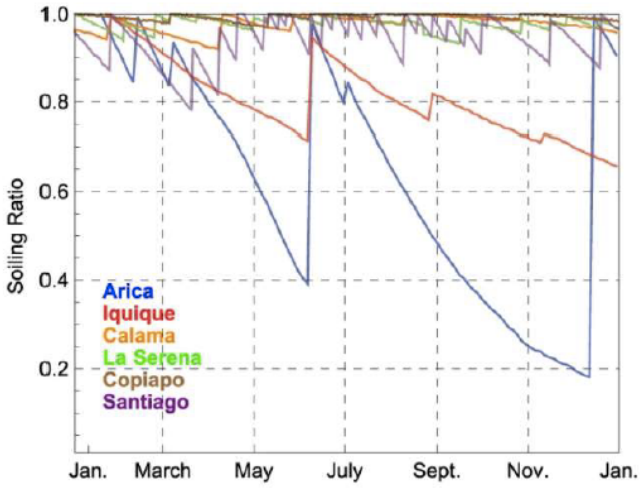

Figura 6. Tasa de ensuciamiento en diferentes plantas solares [19]

La acumulación de tierra u otros elementos en el módulo fotovoltaico puede llevar a una disminución significativa de la energía producida por el módulo fotovoltaico. Por lo tanto, si se realiza un estudio de la planta solar y se observa en los resultados que su rendimiento disminuye con el paso de los años, se debe comprobar si este efecto es debido al ensuciamiento o realmente existe una degradación en los paneles solares. Porque en un estudio de la degradación del campo solar, el efecto de ensuciamiento a largo plazo es ruido añadido a las medidas.

Por eso, este efecto de ensuciamiento se clasifica como determinante en la eficiencia del módulo solar, pero se debe destacar que puede crear confusión a la hora de estudiar la degradación. Ya que puede parecer que la disminución de energía producida es debido a la degradación del campo solar y sin embargo es debido al ensuciamiento del módulo solar.

3.6.5 Degradación

En este proyecto de fin de máster se desea estudiar la degradación del campo solar. Dentro de las plantas solares, las cuales tienen muchos componentes (inversores, cableado, baterías, sensores, etc.) es importante estudiar la degradación del campo solar ya que esta es la parte más costosa de toda la planta.

La planta solar tiene muchos elementos además del campo solar y por lo tanto estudiar la degradación de una planta solar da una visión más general de la degradación ya que engloba otras degradaciones como por ejemplo puede ser la degradación de los sensores, del inversor, etc., las cuales no se tienen en cuenta en la degradación del campo solar. Sin embargo este estudio va a tratar exclusivamente la degradación del campo solar.

Esta degradación se puede apreciar con la disminución del rendimiento de los campos solares. Y esta disminución del rendimiento puede ser detectada mediante un seguimiento de la potencia a lo largo de los años.

Si se observa una disminución de potencia con el paso de los años que no es debida a las condiciones climáticas o a los efectos de ensuciamiento se trata de degradación del campo solar.

El laboratorio de investigación de energías renovables (NREL) ha realizado un estudio empírico utilizando distintas instalaciones y han comprobado que con el paso de los años la mayoría de las instalaciones se degradan [20].

Esta degradación se mide utilizando el coeficiente de tasa de degradación o "*Degradation rate*" que se mide en porcentaje anual (%/año) [19]. Y en ese estudio se indica que la media de degradación anual de los campos solares es de aproximadamente 0.8%/año como se ve en la Figura 7.

Esto parece muy poco, pero las plantas solares trabajan con altas potencias de entorno a los 300 MW y por lo tanto, la perdida anual de potencia puede ser muy significativa y por lo tanto puede suponer una pérdida económica muy importante para la empresa que genera esta energía renovable.

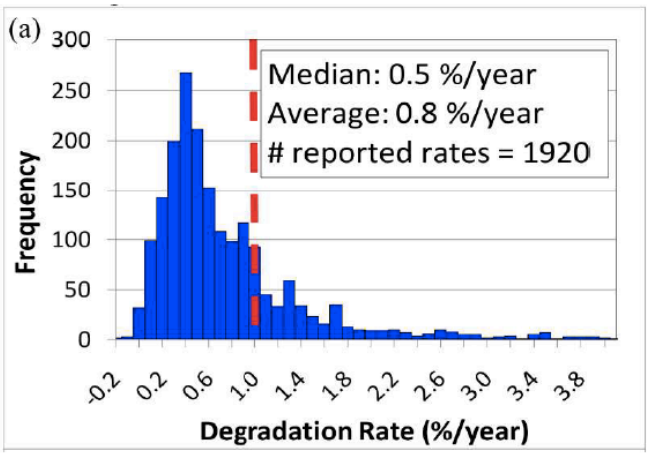

Figura 7. Tasa de degradación en porcentaje anual [20]

Cuando esto se plantea a largo plazo, supone una gran pérdida. Porque la planta solar tiene unos 20-25 años de vida útil y cada año se pierde casi un 1%, y este porcentaje va aumentado (ya que la degradación no es lineal y a medida que pasan los años, la degradación anual aumenta)

Hay algunas plantas que en las que el campo solar se degrada hasta un 3.8% anual, como se puede apreciar en la Figura 7 . Esto no son valores de tasa de degradación comunes, por lo que puede ser debido a módulos que se han salido a la venta mal de fábrica.

Aunque estas degradaciones tan altas no suelen ser lo habitual, como vemos en la Figura 7 el máximo de la gráfica de la distribución se encuentra en 0.4%/año.

Como existen unas garantías concretas que afectan al campo solar, se suelen realizar estudios de cómo se degradan los campos solares comprados para saber si se cumplen las garantías prometidas por el vendedor. Y en caso de que no se cumplan, proceder con la sustitución o devolución de los módulos solares.

Por ejemplo, si se realiza un estudio de la garantía de Trina Solar, se puede observar que esta empresa, que es una de las más punteras en fabricación de módulos fotovoltaicos, garantiza a sus compradores que sus productos no pierden más del 10% de la potencia de salida especificada durante los 10 primeros años. Y no pierden más del 20% de la potencia durante los 15 años siguientes.

NREL en el estudio realizado de la degradación del campo solar ha comprobado que realmente si que existen módulos que no cumplen la garantía. Se puede visualizar en la Figura 8, que entre 2005 y 2010 hay 3 puntos azules que corresponden a 3 tasas de degradación fuera de la garantía del módulo establecida por el fabricante.

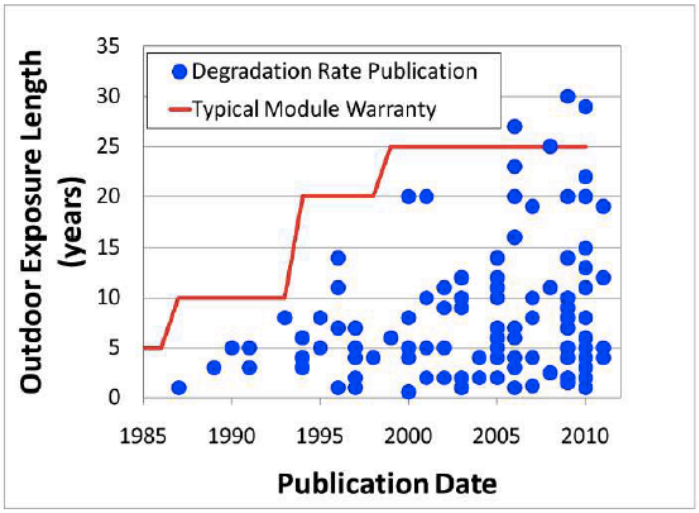

Figura 8. Limites de garantías y tasas de degradación de los módulos [20]

Como estas degradaciones del campo solar suponen mucho dinero perdido para las empresas que realizan instalaciones fotovoltaicas, estas empresas han decidido estudiar cómo se degrada el campo para saber si tienen que reclamar al fabricante en caso de que no cumplan la garantía prometida.

A continuación se van estudiar los efectos que causan la degradación de campo solar.

3.6.5.1 Modos de degradación

Existen diferentes agentes que provocan la degradación de los módulos como puede ser: la luz, la temperatura de las células, los factores ambientales, la fabricación de los módulos, etc. Y se van a estudiar en detalle a continuación, explicando cómo ocurren estas degradaciones y cómo pueden ser detectadas.

FACTORES AMBIENTALES

La humedad es uno de los factores ambientales que degrada el campo solar, ya que la humedad se puede introducir en el módulo y crear oxidación y corrosión en las conexiones eléctricas.

Los cambios de temperatura también influyen en la degradación del campo solar, ya que las células se pueden romper debido a las tensiones que crean los diferentes materiales de la célula al tener distintos coeficientes de dilatación

El granizo también degrada el campo solar al impactar con el módulo, ya que puede romper el vidrio de la célula.

Existen otros agentes no tan comunes pero que si que están en ciertos ambientes como la sal que corroe el metal cuando la planta solar está cerca del mar. O como los gases que crean las industrias, que también crean corrosión si se trata de ácidos.

LID

La degradación inducida por la luz o "LID" (Light Induced Degradation) es la degradación del campo solar producida por la reacción del boro frente a otros elementos químicos contenidos en la célula como el oxígeno, el hierro y el cobre [21]. El problema es que si las células no tienen boro no son igual de eficientes. Ya que el boro es el dopante del silicio que permite que los electrones se muevan.

Así que, cuando se crean las células hay que llegar a un equilibrio entre el boro y los demás químicos.

Este modo de degradación no puede apreciarse a simple vista por lo que para detectarlo se necesita estudiar el rendimiento del módulo. Este tipo de degradación puede alcanzar pérdidas de potencia de hasta un 10% de la potencia inicial.

PID

La degradación inducida por potencial o "*Potential Induced Degradation*" es causada por las corrientes de fuga entre las células del módulo y el resto de componentes del módulo, afectando de manera negativa al rendimiento de estos.

Este efecto puede ocurrir poco después de la instalación de los módulos solares y tiene el problema de no poder detectarse visualmente, sino que se aprecia estudiando el rendimiento.

Este efecto de degradación es causado por las corrientes de fuga entre las células del panel y el resto de componentes del mismo. Se produce un flujo de corriente que fluye entre el marco de aluminio, el vidrio, el tédlar y el EVA, hasta las células del panel. Esto provoca un estrés extremo en las células del panel afectado lo que deriva en una reducción de su rendimiento. [22]

PUNTOS CALIENTES

Los puntos calientes o "*hot spots*" son zonas de la célula con una temperatura alta. Estas zonas solo afectan negativamente en los puntos en los que estas se localizan dentro del módulo solar. Lo que significa que no todo el módulo está degradándose,

sino es por zonas. Estas áreas de alta temperatura disminuye la eficiencia en esa zona en concreto, reduciendo así la potencia de salida de todo el módulo.

Y la degradación de los materiales en esas zonas aumenta debido a las elevadas temperaturas, produciendo así una sensación de degradación de todo el campo solar.

DELAMINACIÓN

La delaminación consiste en la pérdida de adhesión de las diferentes capas que forman un módulo fotovoltaico.

Un módulo solar está formada por varias capas. El módulo se fabrica depositando capas unas encima de otras. Entre estas capas existe una que se denomina EVA [23] que se encarga de actuar como encapsulante, intentando sellar todas las placas para que la célula quede aislada del aire y la humedad [24].

La delaminación ocurre cuando las capas que conforman la célula pierden la adhesión entre ellas. Esta degradación del módulo puede ser debido a que no se han fabricado correctamente las capas o que la capa EVA es de baja calidad.

Se puede detectar visualmente la delaminación del módulo ya que se nota que el módulo cambia de color.

BURBUJAS

Las burbujas son un tipo de degradación muy similar a la delaminación. Ya que la delaminación se produce por la perdida de adhesión en las capas en todo el módulo, y las burbujas se producen solo en una pequeña parte.

Por lo tanto, las burbujas son un aviso de que se va a producir una delaminación. Esta degradación se llama así porque las zonas afectadas tienen una pequeña burbuja, en la que la célula tiene dificultades para disipar el calor y por lo tanto aumenta aún más la degradación del módulo solar.

FALLOS HUMANOS

Durante la instalación de las plantas solares pueden haber roturas de los paneles al llevarse algún golpe o que se ralle el vidrio debido a errores de los técnicos. Estos problemas debido a los fallos humanos que no estaban apenas contemplados pueden causar una degradación del campo solar.

Puede ocurrir solo un tipo de degradación explicada o pueden combinarse varias degradaciones provocando que la degradación del campo solar sea mayor y más rápida, haciendo así que la vida útil del módulo solar disminuya considerablemente.

Por todos estos modos de degradación es necesario cuantificar y predecir cual va a ser esta degradación final del campo solar con el paso de los años.

3.6.5.2 Actuar ante la degradación

Para evitar la degradación se deben tener en cuenta unas pautas antes de comprar el módulo. Y cuando se obtiene degradación en los módulos comprados existen opciones para combatir la degradación.

La degradación del campo solar siempre va a existir y no se puede eliminar su efecto, pero si siguen unos consejos, se puede evitar que el vendedor se aproveche del desconocimiento de garantías del producto.

Ya que todas las placas solares deben tener una garantía del fabricante en la que el fabricante se compromete que los módulos que está vendiendo van a tener un rendimiento mínimo en una cantidad determinada de años.

Antes de comprar los módulos se debe:

- Asegurarse que se trata de un vendedor con varios años en el sector y que fabrica los módulos con materiales de calidad.
- Verificar que ofrece una garantía adecuada

Tras haber comprado los módulos si existe degradación sobre estos, las opciones para solucionarlo son:

- Mantenimiento de la planta solar: limpieza de la planta solar para evitar que la degradación vaya a más.
- Comprobar si los módulos comprados están produciendo la potencia suficiente. Se debe verificar la garantía de rendimiento que prometía el vendedor de los módulos se está cumpliendo. Sino es así se procede con la devolución de los módulos.

En un módulo solar existen dos garantías: la garantía de rendimiento y la de producto. La garantía de rendimiento de un módulo solar suele garantizar que los módulos en los 10 primeros años deben proporcionar al menos el 90% de la potencia nominal. Y un 80% de la potencia nominal a los 25 años. Y la garantía de producto generalmente garantiza que el módulo no va a tener fallos en los primeros 10 o 12 años.

Los módulos monocristalinos tienen una vida útil muy larga ya que ofrecen garantías de hasta 25 años. Y Los policristalinos generalmente tiene menor vida útil que los monocristalinos.

Capítulo 4 - Caso práctico de estudio

Este capítulo presenta un estudio práctico de la degradación de una planta solar real de una empresa. Y además, se pretende realizar un estudio de las predicciones de potencia a través de sensores para conocer a priori el rendimiento que se obtendría si se colocase realmente una planta solar en la zona en la que se han ubicado los sensores.

Para ambos estudios, es necesario un tratamiento y análisis de datos, y para ello se va a usar Python. En concreto, el entorno de trabajo en el que se va a realizar este estudio es Jupyter Notebook.

A continuación se va a realizar una introducción de la estructura de la planta solar analizada; también de los datos que se han medido de esta planta para saber la cantidad de datos que hay, la distribución de estos y su nomenclatura.

Por lo tanto este capítulo va a estar dividido en: la estructura de la planta analizada, preprocesado de los datos, y visualización de los datos.

4.1 Estructura de la planta analizada

Los datos que se van a analizar en este estudio son privados. Estos datos pertenecen a una empresa que se dedica a la energía solar. Esta empresa nos ha cedido los datos y en este estudio no se va a publicar el nombre de la empresa, ni el nombre de la planta solar, al igual que tampoco se van a mostrar los datos analizados por compromiso de confidencialidad.

Como ya se ha explicado anteriormente, el trabajo del inversor es convertir la corriente continua que procede de las placas solares en corriente alterna. Independientemente de qué se haga con la corriente alterna que sale del inversor, ya que puede ir a baterías o directamente a la red eléctrica, en este estudio se va a analizar la corriente continua. Es decir, se estudia la corriente que va desde el campo solar hasta el inversor. Y del inversor hacía fuera no se va a tener en cuenta en este estudio. Por lo tanto, se destaca que cuando se nombra potencia en este estudio, se trata de la potencia de continua, que es la multiplicación de tensión continua y

corriente continua. Y aunque no se especifique, la potencia es continua (DC) y no alterna (AC). Esto es debido a que este estudio en cuestión va a tratar de la degradación del campo solar, y no de la degradación de la planta solar entera, que sería demasiado complejo y habría que tener en cuenta muchos más componentes de la planta que aportan degradación.

También se debe conocer cómo es la estructura del campo solar del que se van a analizar los datos. Y por ello se va a explicar cómo se organizan los elementos más simples para formar estructuras mayores.

Se procede a comenzar por la unidad más básica y pequeña, que es la célula fotovoltaica. La potencia que proporciona una célula es muy pequeña (alrededor de 1 W o 2 W). Por lo que varias células se agrupan para proporcionar mayor potencia. En la Figura 9 se observa cómo estas células conectadas en serie y paralelo entre sí, forman un grupo de células que se denomina panel o módulo solar. Los paneles solares más habituales tienen 60 células y suelen proporcionar una potencia entre 230 y 330 W.

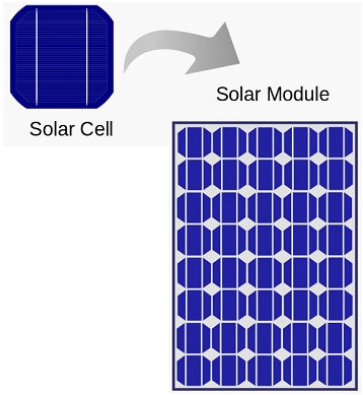

Figura 9. Módulo solar compuesto por células solares

Las células conectadas en serie permiten aumentar la tensión final de la célula equivalente. Y las células conectadas en paralelo permiten aumentar la intensidad final del conjunto. Entonces como aumenta la tensión y la corriente del conjunto equivalente, la potencia del conjunto también aumenta al ser producto de estas dos.

Aunque se haya aumentado la potencia agrupando las células, la potencia que proporciona un módulo solar aún es pequeña. Entonces, para aumentar la potencia se

conectan primero varios módulos solares en serie y posteriormente en paralelo para así conseguir mayor potencia en el conjunto.

Por lo tanto, los módulos solares se conectan en serie para formar un panel o *string* solar como se visualiza en la Figura 10.

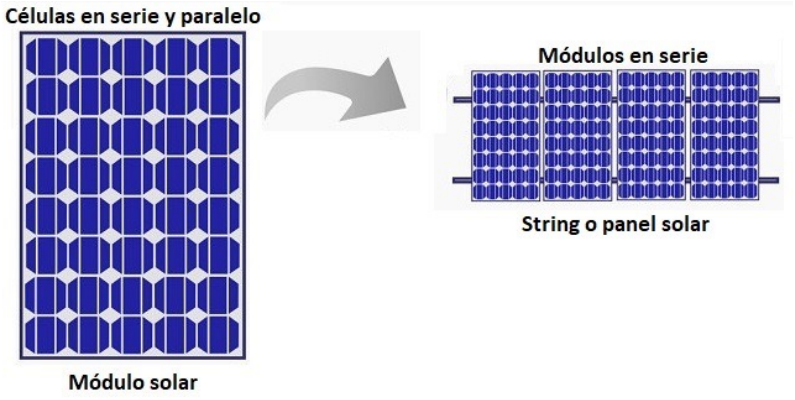

Figura 10. Módulos en serie forman un panel solar

Una vez que se han creado varios *strings* conectando los módulos en serie, se deben conectar estos paneles o *strings* en paralelo para formar un *array* o campo solar. Tal y como se indica en la Figura 11.

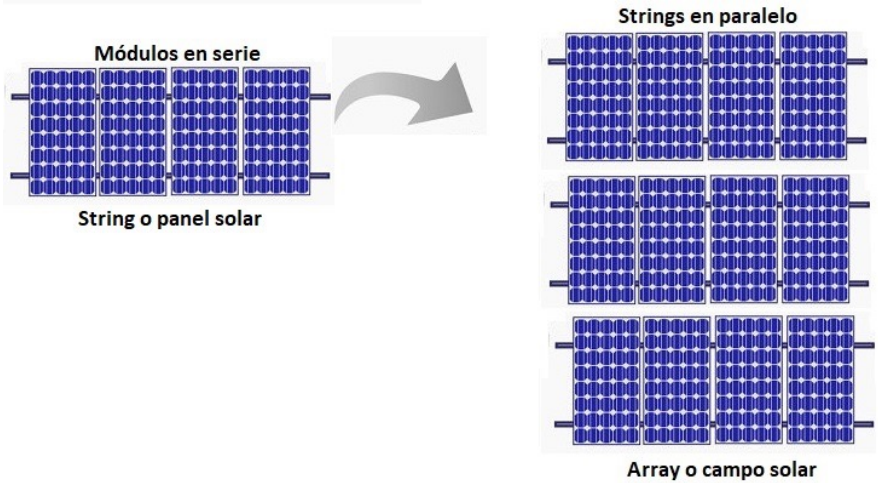

Figura 11. Agrupar varios strings en paralelo para formar un array o campo solar

La estructura de cómo se conectan eléctricamente los diferentes elementos solares para formar conjuntos mayores y obtener mayor potencia es común a todas las plantas solares. Al igual que la nomenclatura de estos grupos, aunque puede variar dependiendo del país.

A continuación se detalla la cantidad de módulos, paneles y campos solares que dispone la planta solar que se va a estudiar. Esto es necesario para que posteriormente se entiendan correctamente los datos que han sido proporcionados para este estudio de la degradación solar. Cabe destacar que esta distribución o estructura que a continuación se va a explicar no es común para todas las plantas solares, sino que se trata de la distribución de la planta a estudiar.

Es interesante realizar cálculos previos de la potencia que el campo solar va a enviar al inversor ya que si es mayor que la potencia el inversor soporta, el inversor va a limitar la potencia y esta se va a perder. Por lo tanto no todos los *strings* o paneles tienen que tener la misma cantidad de módulos, sino que la cantidad de módulos en cada panel puede variar.

En vez de desaprovechar la energía que se estaría perdiendo en el inversor, se intentan hacer cálculos del número de módulos necesarios en los *strings* para "cuadrar" la potencia que se va a enviar al inversor.

En esta planta existen *strings* de 14 y 16 módulos. Es decir, se conectan 14 o 16 módulos en serie para formar un único *string* o panel solar, como se índica en la Figura 12. Y posteriormente se combinan estos *strings* en paralelo para formar un *array* o campo solar. En concreto se conectan 3 *strings* en paralelo como se observa en la Figura 12.

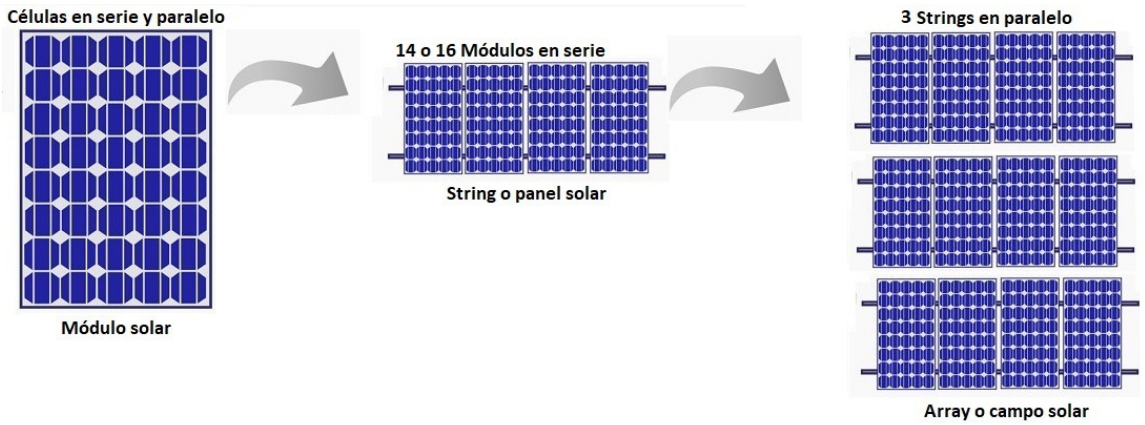

Figura 12. Distribución planta solar a estudiar

Ahora que ya se ha mostrado la estructura que tiene el campo solar, se muestra la estructura completa de la planta solar en la Figura 13. En una planta solar suelen haber varios campos solares. En este caso, la planta dispone de 4 campos solares como se observa en la siguiente figura. Estos campos solares van conectados al inversor. El inversor es tratado como una caja negra que transforma la potencia de continua a alterna. Aunque realmente ese inversor es un conjunto de 4 inversores. En el que cada uno de esos 4 inversores va a recibir la potencia de solo un campo solar. Es decir, como si cada campo solar estuviese conectado a un inversor. Por lo tanto, los campos solares se pueden tratar como independientes. Y se puede calcular la degradación de cada campo solar por separado, que es lo que se va a realizar en este capítulo. Como se ha mencionado anteriormente, en este estudio no se va a calcular la degradación de la planta solar ya que habría que tener en cuenta muchas degradaciones de otros elementos del sistema. Por lo que no se va a calcular una degradación de los 4 campos solares como un conjunto, sino la degradación de estos 4 campos por separado ya que los campos son independientes, y por lo tanto la degradación de cada campo también lo es.

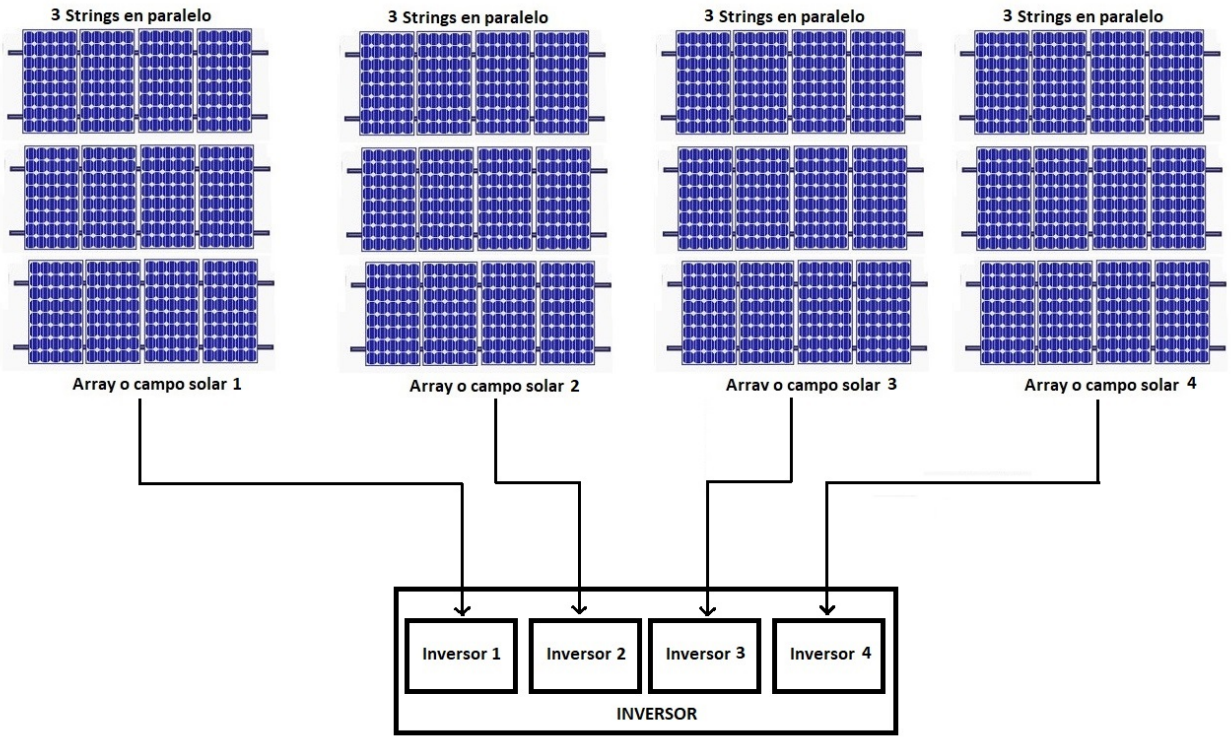

Figura 13. Esquema general de los campos solares e inversores

En la Figura 14 se puede apreciar el número de módulos solares que contiene cada panel solar. Además esta figura da una visión global de cómo es la estructura completa de la planta solar.

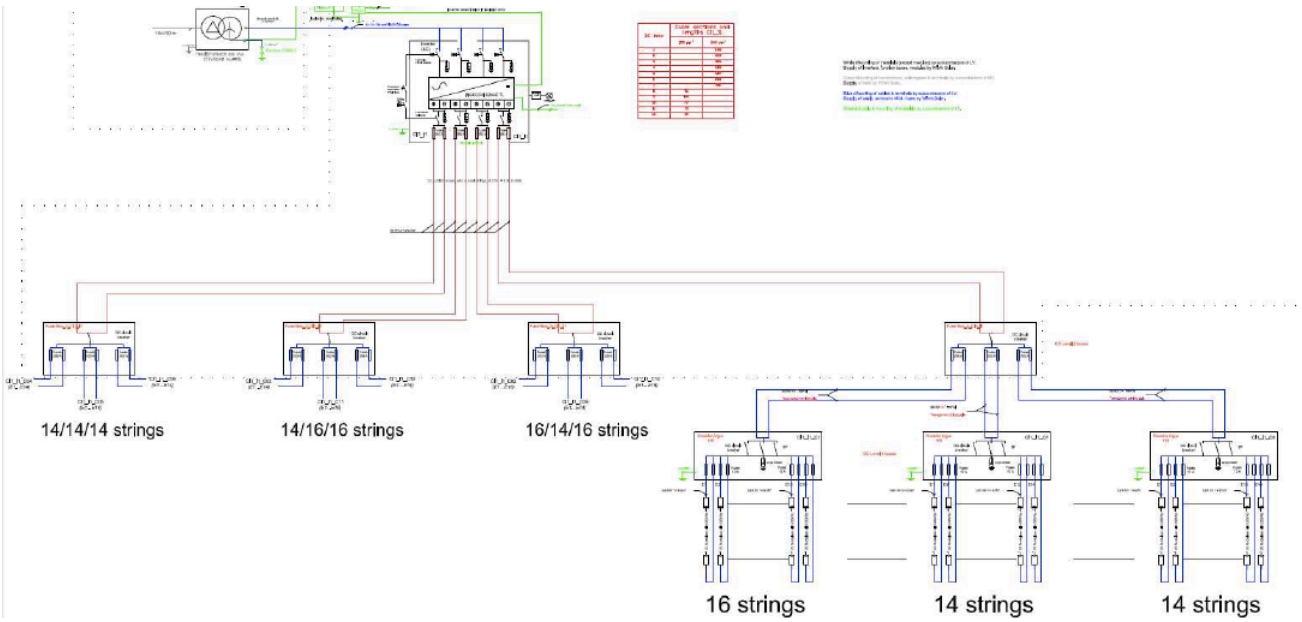

Figura 14. Esquema de la estructura general de la planta solar

4.2 Preprocesado

Los datos que la empresa ha proporcionado para el estudio de la degradación van desde el año 2012 hasta el 2020 inclusive. Por lo tanto, se dispone de 9 archivos en formato csv.

Cada archivo tiene el mismo número de columnas y la misma nomenclatura. Por lo que se pueden cargar los 9 archivos en Python, cada uno como un *dataframe*.

Un *dataframe* es una estructura tabular ofrecida por la biblioteca *Pandas* de Python para almacenar y manipular datos.

Una vez importados los 9 *dataframes*, se van a unir todos en único *dataframe* que contenga los datos de todos los años.

El resultado de unir todos los años en un mismo *dataframe*, permite calcular las filas que este tiene. Se trata de 315621 filas y 17 columnas. Es decir un total de 5365557 valores.

Como se puede apreciar, hay una gran cantidad de datos. Esto es debido a que el muestreo de los datos es cada 15 minutos. Por lo tanto, se obtienen 17 valores cada 15 minutos durante 9 años.

Las columnas del *dataframe* son las siguientes: 'Fecha', 'Icc_11', 'Vcc_11', 'Pac_11', 'Icc_12', 'Vcc_12', 'Pac_12', 'Icc_13', 'Vcc_13', 'Pca_13', 'Icc_14', 'Vcc_14', 'Pca_14', 'RadPanel_1' 'RadPanel_2', 'TempMod_1' y 'TempMod_2'.

Para las magnitudes mencionadas se utilizan las unidades básicas del sistema internacional. La corriente se mide en amperios, la tensión en voltios, la potencia en vatios. La radiación en vatios partido metro cuadrado y la temperatura en grados Celsius. Y en términos de energía, como las muestras son cada 15 minutos, se mide en vatios cada 15 minutos; no vatios cada hora, como es habitual.

A continuación se va a explicar el contenido de cada columna.

- La columna ‘Fecha’ contiene la fecha (mes/día/año hora : minutos) en la que se miden las muestras. Como ya se ha mencionado se toma una muestra cada 15 minutos.

- Las columnas: ‘RadPanel_1' y 'RadPanel_2' contienen las medidas de 2 piranómetros (sensores de radiación) situados en el plano del panel. Por lo tanto la radiación captada de ambos sensores en un mismo campo solar debería ser similar.
- Las columnas: 'TempMod_1' y 'TempMod_2' contienen las medidas de 2 sensores de temperatura colocados en el plano del panel solar. Estas dos temperaturas también deben ser similares entre sí, aunque la temperatura es más susceptible al viento. Sin embargo, tanto la radiación como la temperatura no tiene por que ser tan similar entre diferentes campos solares. Aunque se encuentre los 4 campos solares en la misma ubicación, y por lo tanto con las mismas condiciones meteorológicas, quizás se encuentra alguna nube más sobre un campo solar que de otro. Y al estar los campos solares distanciados, aunque parezca que se encuentran exactamente en las mismas condiciones del entorno pueden existir diferencias pequeñas de temperatura.
- Las columnas: 'Icc_11', 'Vcc_11', 'Icc_12', 'Vcc_12', 'Icc_13', 'Vcc_13', 'Icc_14' y 'Vcc_14' son las columnas de corriente y tensión continua de los 4 campos solares. El número del final es el que indica el número del campo solar. Por eso van del 1 al 4.
- Por otra parte están las columnas: 'Pac_11', 'Pac_12', 'Pca_13' y 'Pca_14', que son las potencias de alterna de los 4 campos solares. Las cuales no son de interés en este estudio, por lo que van a ser eliminadas.
- En cambio, las potencias que si que son de interés en este estudio son las de continua, las cuales no están incluidas en los datos originales, pero pueden ser calculadas con los datos existentes. Ya que se puede calcular la potencia continua como producto de la tensión continua y la corriente continua: Pcc_11= Vcc_11 x Icc_11. Y lo mismo ocurre para el resto de campos solares. Una vez calculada la columna de potencia en continua para los 4 módulos: ‘Pcc_11’, ‘Pcc_12’, ‘Pcc_13’, ‘Pcc_14’ a través de la tensión y la corriente en continúa, se debe añadir al *dataframe* . Por lo que ahora existen 4 nuevas columnas de potencia continua. Pero por otro lado, se van a eliminar las columnas de tensión y corriente que ya se han utilizado para el cálculo y que no van a ser de interés.

Como se ha mencionado, en el *dataframe* resultante hay 4 columnas de potencia distintas, una por cada módulo. El siguiente paso consiste en dividir el *dataframe* en 4 *dataframes*, cada uno con los datos correspondientes a cada módulo. Como los campos solares son independientes, es recomendable que cada campo solar sea un *dataframe* diferente.

Así que, el conjunto de datos obtenido tras añadir las nuevas columnas y suprimir las que sobran se divide en 4 conjuntos de datos, uno por cada módulo solar; y van a ser tratados por separado.

Una vez realizada la división de todos los datos en 4 *dataframes*. Cada uno de estos contiene las siguientes 6 columnas: 'Fecha', 'Pcc_1X' (columna calculada), 'RadPanel_1', 'RadPanel_2', 'TempMod_1' y 'TempMod_2'.

Y las filas siguen siendo las mismas para los 4 *dataframes* ya que no han sido modificadas.

Tras la división, cualquier cambio o análisis de datos que se quiere realizar sobre el *dataframe* tiene que realizarse 4 veces (una por cada campo solar). Para no repetir el mismo código para cada *dataframe*, se crean funciones con un código parametrizado que permite llamarlo 4 veces y pasarle como parámetro el *dataframe* deseado.

4.2.1 Pasos del preprocesado

Hasta ahora, se ha explicado la primera parte de la fase de preprocesado. La cual ha consistido en unir todos los años, eliminar columnas innecesarias y dividir las columnas en los campos solares correspondientes, como se observa en la Figura 15.

Figura 15. Esquema fase de preprocesado 1

Ahora que ya se tienen los datos divididos por campos solares y que se pueden tratar independientemente se va a explicar la segunda parte de la fase de preprocesado. El diagrama de flujo de la Figura 16. contiene todos los pasos de la segunda parte de la fase de preprocesado.

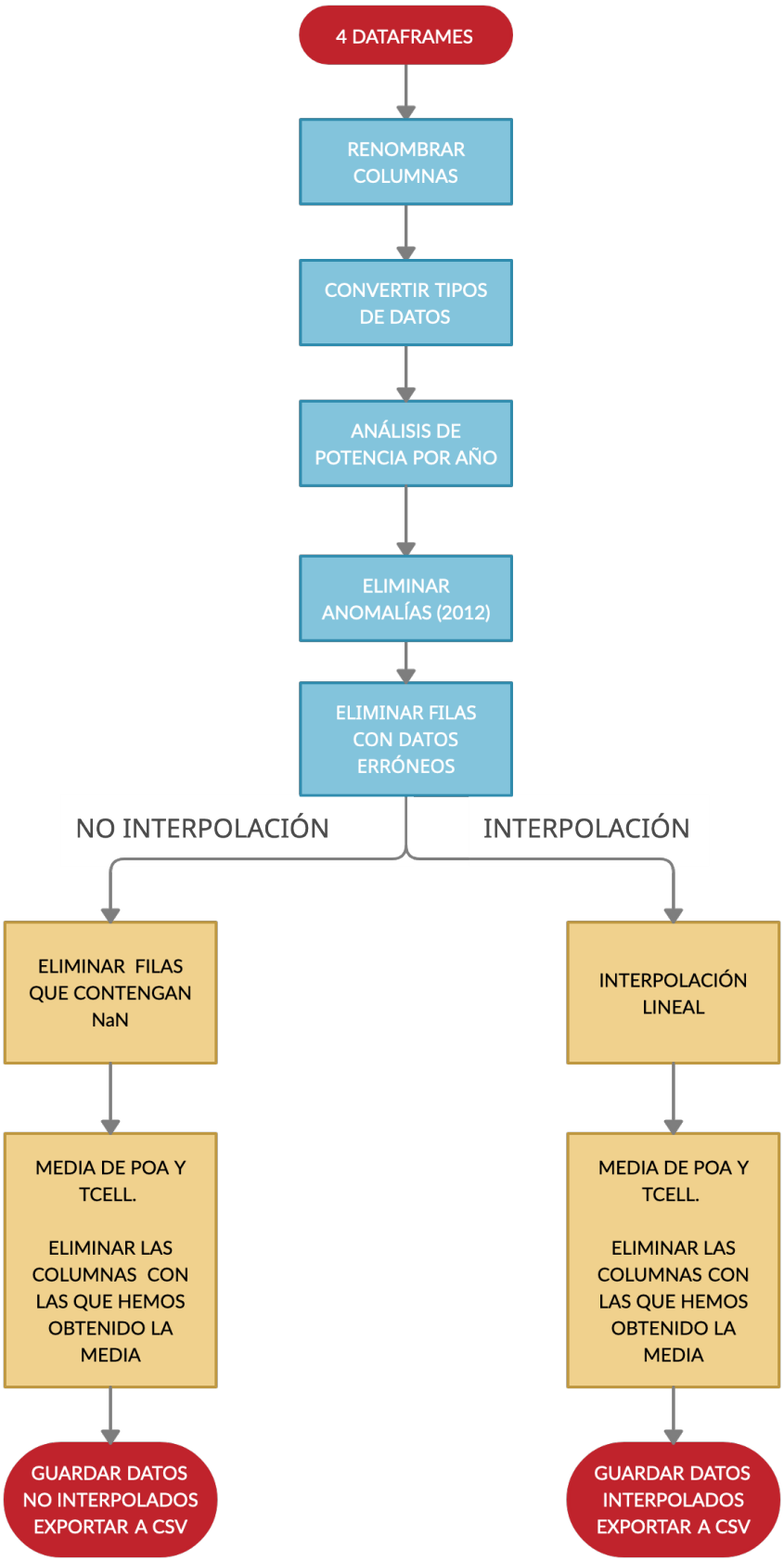

Figura 16. Esquema fase de preprocesado 2

Estos pasos del preprocesado vistos en la Figura 16 se van a explicar en detalle en los siguientes subsecciones.

4.2.2 Renombrar columnas

Para simplificar la nomenclatura de las columnas se renombran en todos los campos solares: "Timestamp"(Fecha), "power"(Pcc_1X), "poa"(RadPanel_1), "poa2"(RadPanel_2), "Tcell" (TempMod_1) y "Tcell2" (TempMod_2). Siendo la palabra entre paréntesis el nombre anterior de la columna.

4.2.3 Convertir tipos de datos

Si se comprueba el tipo de las columnas utilizando *dtype*, se puede observar como algunas columnas son de tipo numérico *float* : *dtype('float64')*, sin embargo hay otras que son del tipo objeto de Python: *dtype('O')*. Por lo que se ha decidido convertir todas las columnas a numérico utilizando la función de *Pandas* "*to_numeric*", menos para la columna "Timestamp", la cual va a ser transformada a *Datetime* usando la función de *Pandas* "*to_datetime*".

Cuando se ha realizado esta conversión de tipos para obtener en todas las columnas valores numéricos, a la función "*to_numeric*" se le ha indicado en el parámetro de errores que sea "coerce". Lo que significa que si en los datos existe algún valor que sea inválido, que no puede transformarse a numérico; lo establezca como *NaN*. El valor *NaN* significa "*Not a Number*" y es una constante que indica que el valor no es válido.

4.2.4 Análisis de potencia por año

Seguidamente, se va a visualizar la potencia media de cada año mediante una gráfica, para tener una idea de cómo son los datos. Se podría pensar en realizar esta gráfica para poder visualizar si existe degradación con el paso de los años. Ya que la

degradación solar, se ve afectada por el rendimiento del campo solar y por tanto por la potencia. Sin embargo no es tan sencillo como esto. Ya que un año se ha podido tener menos radiación que en otro año, por haber sido un año más nublado. Y por lo tanto al tener menos radiación ese año, se obtiene menos potencia que en el año anterior y se podría afirmar equivocadamente que es debido a la degradación del campo solar. Sin embargo es debido al clima.

Esta comprobación se hace para verificar si existe algún año que tenga alguna anomalía, valores atípicos o "*outliers*". Ya pueden ser valores de potencia nulos que pueden ser causados por un sensor estropeado, apagados temporales, ensuciamiento de los módulos, valores inválidos (*NaN*), etc.

Tras haber pasado la potencia de vatios a kilovatios. Se calculan las medias de la potencia para cada año y para cada uno de los 4 campos solares.

Por lo que la Figura 17 corresponde con el df1 que contiene los datos del campo solar 1.

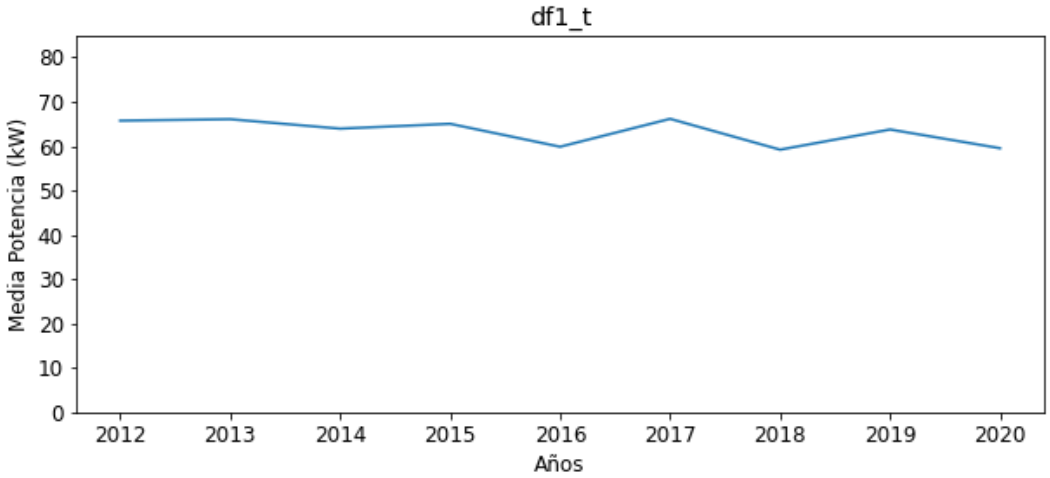

Figura 17. Media de potencia en el campo solar 1

Y la Figura 18 corresponde con el df2 que contiene los datos del campo solar 2. Por lo que en esta figura se está viendo la evolución de la potencia a lo largo de los años solo del campo solar 2. Y así con el resto: Figura 19 y Figura 20. El df3 corresponde con el campo solar 3 y el df4 con el campo solar 4.

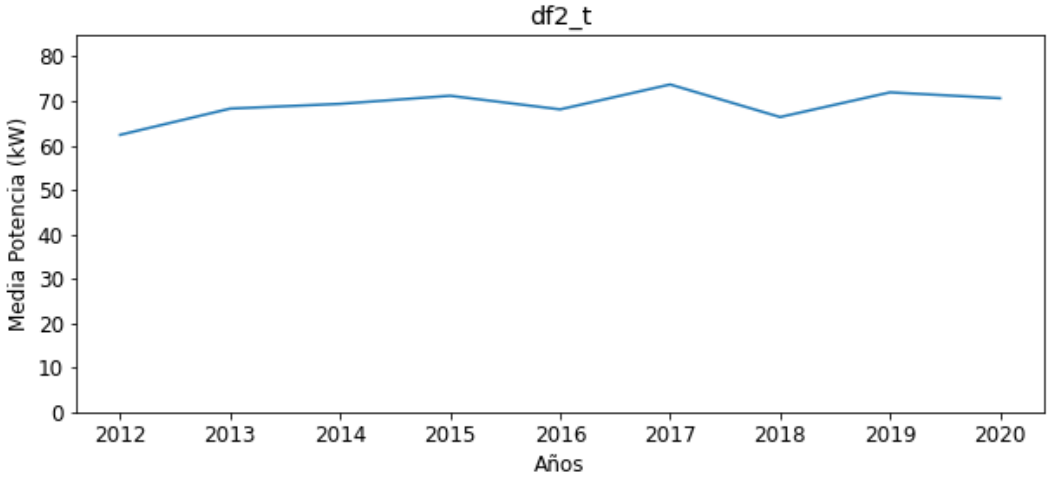

Figura 18. Media de potencia en el campo solar 2

Como se puede apreciar, los 4 campos solares a pesar de ser independientes y poder realizar cálculos sobre ellos de forma modular, tienen una apariencia muy similar. Y todos ellos rondan los mismos valores cuando se hace una media. Al estar los campos solares ubicados próximos unos a otros las condiciones del entorno van a ser parecidas, aunque nunca iguales.

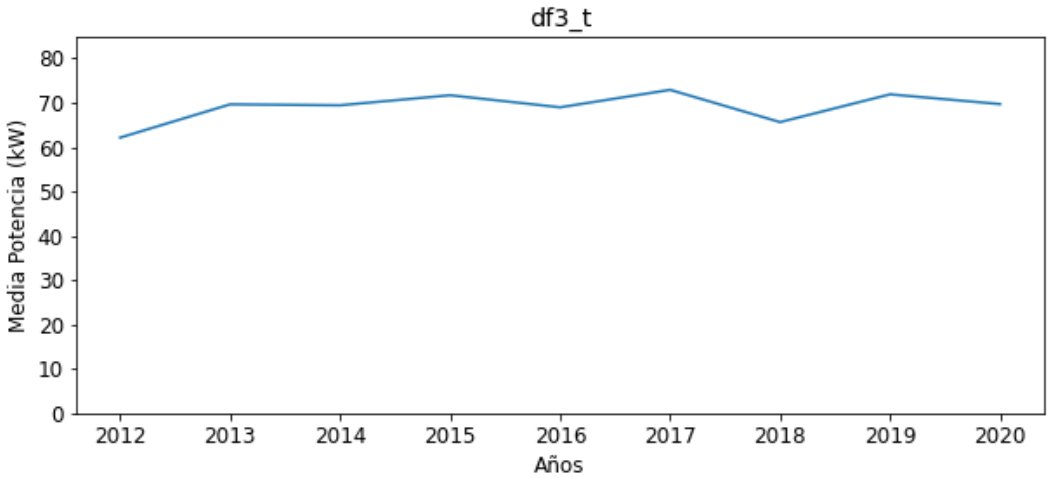

Figura 19. Media de potencia en el campo solar 3

Si se observan la Figura 17, la Figura 18, la Figura 19 y la Figura 20 se puede comprobar que las medias de las potencias anuales rondan entorno los 60 y los 70 kilovatios.

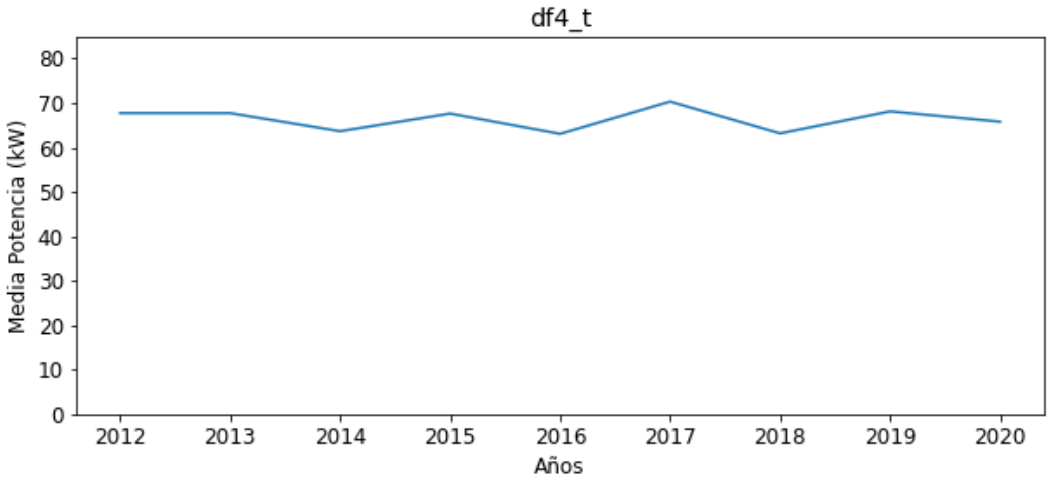

Figura 20. Media de potencia en el campo solar 4

4.2.5 Eliminar anomalías

Si estas gráficas se observasen con mayor detalle se podrían apreciar estas anomalías que se mencionaban anteriormente. Para ello, se representa la misma gráfica (media de la potencia anual) en cada campo solar, pero dejando la potencia en W en vez de kW para observar que es lo que está ocurriendo con más detalle.

Y no se observa nada para el campo solar 1 y 4. Sin embargo en el campo solar 2 y 3, que corresponde cada uno con la Figura 21 y la Figura 22 respectivamente; sí que se observa que el año 2012 tiene unos valores muy bajos de potencia.

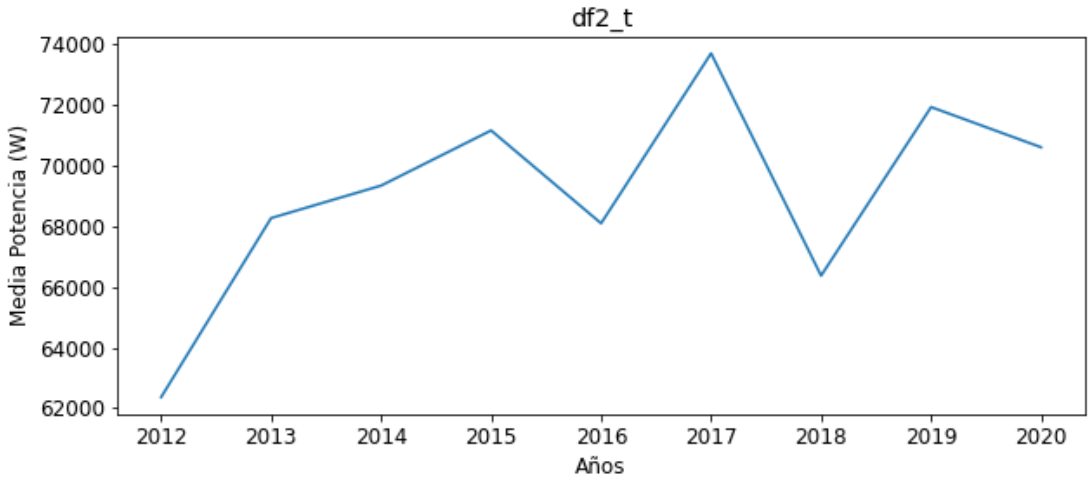

Figura 21. Media de la potencia del campo solar 2 en detalle

No es habitual que este valor se desvíe tanto de otros años. Y este efecto anómalo se observa en dos campos solares. Esta anomalía puede dar a entender que el primer año que se capturaron los datos de la planta aún no estaba en su correcto funcionamiento. Por lo que este año no debería ser tratado en este estudio de la degradación ya que estos datos no son fiables y podría afectar negativamente al estudio, por lo que se ha decidido eliminar el año 2012. Como los campos son independientes no tendría por qué eliminarse en los campos 1 y 4, donde al parecer no ocurre este efecto. Pero para seguir una misma directriz para toda la planta, se decide que los datos del año 2012 deben ser eliminados en todos los campos solares. El nuevo *dataframe* con el año 2012 eliminado se nombra como "dfX_acortado".

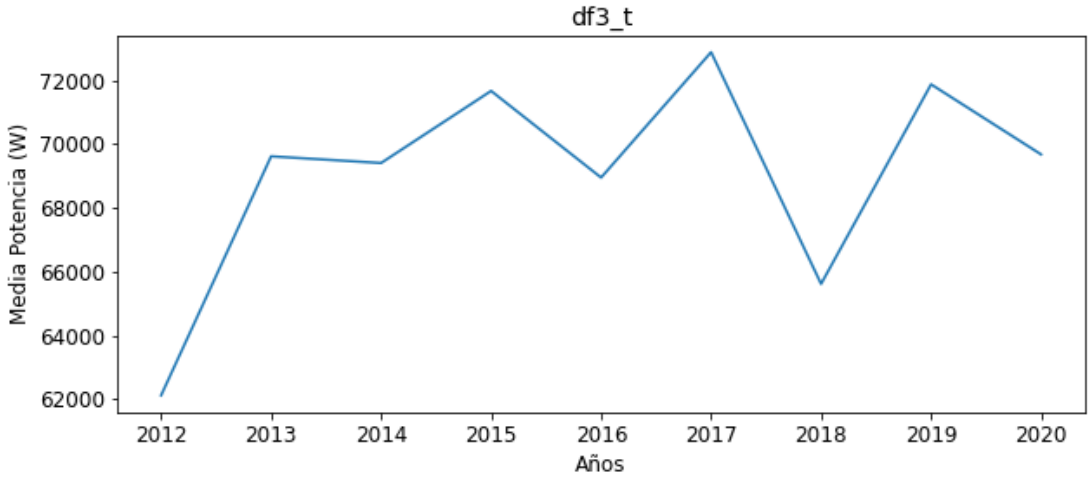

Figura 22. Media de la potencia del campo solar 3 en detalle

En los datos iniciales había 315621 filas y ahora que se ha eliminado el año 2012 solo quedan 280489 filas. Lo que significa que se han eliminado 35132 filas.

Por lo tanto cada *dataframe* que hace referencia a un campo solar está compuesto por 280489 filas y 6 columnas ('Timestamp', 'power', 'poa', 'poa2', 'Tcell' y 'Tcell2)

4.2.6 Eliminar filas con datos erróneos

A continuación sería conveniente hacer un análisis más exhaustivo de la potencia por si es necesario realizar una limpieza de los datos. Para ello se debe comprobar si contiene valores que no aportan nada al estudio, o incluso valores que añaden ruido al estudio; y en caso de que existan, eliminarlos.

Los valores de potencia igual a cero, significan que realmente no hay potencia; ya sea por falta de radiación solar o avería de algún componente, o parada del equipo por algún motivo (mantenimiento u orden del operador de red). Por lo tanto las filas que contengan potencia nula, van a ser eliminadas.

Se han detectado 16032 filas con potencia igual a cero en el campo solar 1. En el campo solar 2: 14204 filas con potencia igual a cero. En el campo solar 3: 15103 filas con potencia igual a cero. Y en el campo solar 4: 16117 filas con potencia igual a cero.

Todos los campos rondan los mismos valores de potencia nulos; entre unos 14 mil y 16 mil.

4.2.7 Posibles soluciones frente a valores inválidos

En esta fase de preprocesado de los datos, de nuevo se encuentra otro problema. Se dijo con anterioridad que los valores inválidos que no se pudieran transformar a numéricos, iban a tener el valor de *NaN*. Pues estos valores realmente no aportan nada aprovechable al estudio y cuanto más limpios se dejen los datos que se van a analizar, más simple será la siguientes fases.

¿Qué se hacen con estos valores *NaN*? Pues antes de nada se debe ser la cantidad que se tiene de estos. En el campo solar 1 han quedado 264457 filas tras haber eliminado 2012 y haber eliminado las filas que contenían valores nulos de potencia. Y en estas filas se han encontrado 152278 filas con valores *NaN*. Con que 1 de los 6 valores de una fila sea *NaN*, ya se considera como que fila que tiene un valor inválido y debe ser estudiada. Y en los demás campos solares se obtienen resultados parecidos en cuanto a filas con algún *NaN*.

Se trata de muchas filas que tienen valores *NaN*, más de la mitad de las filas. En concreto un 57.58% de las filas actuales contienen mínimo un *NaN*.

Al ser muchos datos los que están implicados, una de las opciones es intentar "arreglar" estos datos. Y la otra opción, que es la sencilla es eliminar esas filas que contengan algún valor inválido.

La primera opción que se ha mencionado consiste en interpolar los valores inválidos para seguir manteniendo el mismo número de filas. Es una lástima eliminar una fila entera que puede tener válidos, por un solo valor inválido. Porque siendo tantos datos los que están involucrados, si se eliminan (la opción más simple) puede que se estén desaprovechando muchos datos los cuales podrían servir para el estudio de la degradación.

Y la segunda opción expuesta, que es la sencilla; simplemente consta de eliminar datos. Esta opción se ha descrito como "No interpolación". Para aclarar que son dos alternativas distintas y por lo tanto dos caminos paralelos, como se puede observar en la Figura 23.

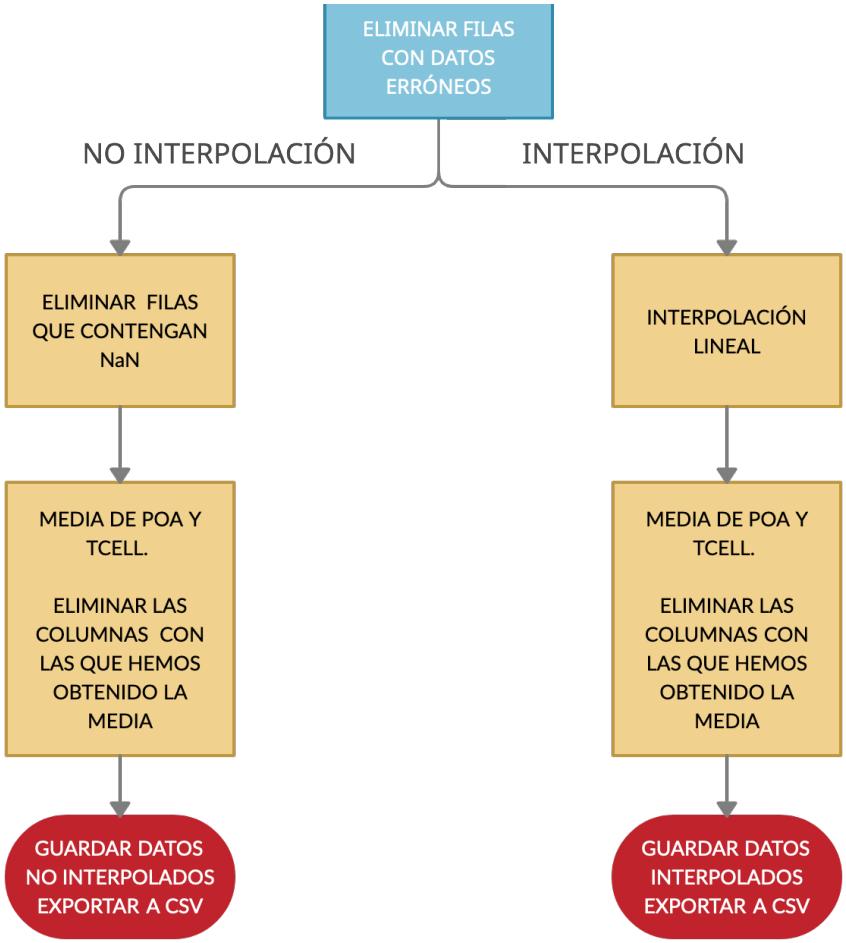

Figura 23. Soluciones frente a valores inválidos

A continuación se van a explicar las dos opciones ya introducidas, que existen para solucionar el problema de los valores inválidos en mayor detalle.

4.2.8 Interpolación

La interpolación lineal es una técnica para estimar los valores que toma una función en un intervalo, del cual se conocen sus valores en los extremos. Para estimar uno o varios valores se utiliza la aproximación a la función por medio de una recta. En definitiva la interpolación trata de deducir un valor o varios valores entre dos valores que sí que están definidos. Lo que se debe entender, es que la interpolación es una aproximación y no es un valor medido. Así que si se comparase el valor medido y la aproximación se vería que la interpolación ha introducido un pequeño error respecto al valor real.

La interpolación más conocida es la interpolación lineal. Consiste en trazar una línea recta que pasa por dos puntos conocidos y debe calcular los valores intermedios según la recta.

Se van a interpolar las siguientes 5 columnas: 'power', 'poa', 'poa2', 'Tcell' y 'Tcell2'. Ya que se ha comprobado que la columna de 'Timestamp' no dispone de ningún valor inválido.

Cuando se encuentra una secuencia de "n" valores inválidos consecutivos. Se debe encontrar el valor anterior válido, que en este estudio se ha definido como "a", también se debe identificar el valor siguiente, que en este estudio se ha definido como "s". Y la posición del valor inválido que se va a sustituir se define como "i". Por lo que si hay "n" valores inválidos consecutivos, "i" va desde 1 hasta "n" (i=1,2,…, n).

El valor interpolado que se debe sustituir en la posición "i" sigue la siguiente formula:

Valor interpolado = $a + i \times \frac{s-a}{n+1}$ con i=1…n

Esta formula también sirve también aunque no se encuentren varios valores consecutivos, es decir se puede aplicar la formula si solo se tiene un *NaN*; donde "n" será igual a 1. Y la posición "i" solo tomará el valor de 1.

A continuación se va a explicar un ejemplo de interpolación para entenderlo mejor.

En la Figura 24 se puede contemplar unos valores *NaN* consecutivos. En concreto 3 valores *NaN* consecutivos. Por lo que "n" igual a 3.

También se puede apreciar que si representamos los datos, existen valores inválidos que no son incluidos en el gráfico. Esos huecos en la recta son debido a los 3 valores *NaN* que existen.

Se va a utilizar la formula mencionada anteriormente para realizar la interpolación. La tabla muestra que el valor anterior al primer *NaN* es el 7, por lo que "a" es igual a 7. Y el siguiente valor a los *Nan* es 15, por lo que "s" es igual a 15. Las posiciones 322, 323 y 324 corresponden con "i" igual a 1, "i" igual a 2, "i" igual a 3 respectivamente.

Figura 24. Ejemplo interpolación parte 1

Con todos estos datos, ya se pueden calcular los valores interpolados para las posiciones 322, 323 y 324.

n=3						
a=7						
i=1	valor interpolado -->	a + i * (s-a)/(n+1) -->	7 + 1 * (15-7)/(3+1) -->	7 + 1 * 8/4 -->	7+2 -->	9
i=2	valor interpolado -->	a + i * (s-a)/(n+1) -->	7 + 2 * (15-7)/(3+1) -->	7 + 2 * 8/4 -->	7+4 -->	11
i=3	valor interpolado -->	a + i * (s-a)/(n+1) -->	7 + 3 * (15-7)/(3+1) -->	7 + 3 * 8/4 -->	7+6 -->	13
s=15						

Figura 25. Ejemplo interpolación parte 2

Para cada posición "i", que significa posición con valor *NaN*; se va a tener un resultado distinto, ya que la "i" en la formula de interpolación lineal es el parámetro que va cambiando. El resto de parámetros (n, a, s) se mantienen igual.

En la Figura 25, para i=1 se obtiene el valor interpolado=9, para i=2 se obtiene el valor interpolado=11, para i=3 se obtiene el valor interpolado=13

Estos valores interpolados se han calculado para sustituirlos por los valores incorrectos, haciendo una predicción lineal. Así pues si se cambian los valores *NaN* por los valores calculados: 9,11 y 13 y de nuevo se muestra la gráfica de estos datos, se puede ver que los datos que antes estaban como huecos ahora están interpolados.

Esta interpolación es lineal y como se puede apreciar en la Figura 26 los valores que han sido interpolados siguen la misma línea recta que los valores ya existentes, como era de esperar.

			POSICION	VALOR
			320	5
			321	7
9	-->		322	9
11	-->		323	11
13	-->		324	13
			325	15
			326	17

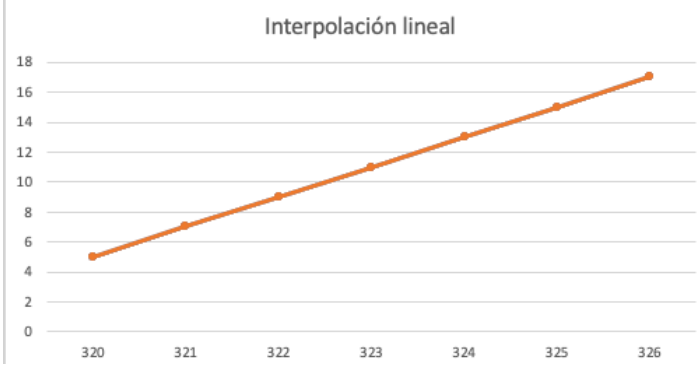

Figura 26. Ejemplo interpolación parte 3

Sin embargo, no todas las situaciones son tan simples como la que se acaba de explicar, ya que también hay que tener en cuenta otros casos más especiales.

Puede ocurrir que los datos *NaN* no se encuentren entre 2 valores válidos. Esto puede ocurrir en 2 casos especiales.

- Los datos empiezan con un valor *NaN*. Es decir el primer dato ya es un valor inválido y por lo tanto no tiene un valor anterior que sirva para realizar la interpolación. También puede ser que no solo sea un valor el del principio sino que existan varios *NaN* al principio consecutivos.
- Los datos terminan con un valor *NaN*. Es decir, el último valor es un data incorrecto y por lo tanto no tiene un valor siguiente que sirva para realizar la interpolación. También puede ser que no solo sea un valor el del final sino que existan varios *NaN* al final consecutivos.

En el caso de que el primer valor de los datos sea uno o varios *NaN*, se selecciona el siguiente valor no nulo y se sustituye en las posiciones que sean *NaN*. Y parecido ocurre cuando los datos finalicen con uno o varios *NaN*. En este caso se busca el último valor válido antes de el/los valor/es *NaN*, se selecciona este valor y se sustituye en todas las posiciones que sean *NaN*.

Un ejemplo del primer caso sería el siguiente.

POSICION	VALOR		POSICION	VALOR
320	NaN		320	13
321	NaN		321	13
322	NaN	--INTERPOLACIÓN -->	322	13
323	NaN		323	13
324	13		324	13
325	15		325	15
326	17		326	17

Figura 27. Ejemplo valores NaN al principio

En la Figura 27 se puede observar como existen 4 valores *NaN* al principio y no existe un valor anterior. Por lo que se elige el valor siguiente válido para sustituir los valores *NaN*, que en este caso es el 13.

POSICION	VALOR			POSICION	VALOR
320	5			320	5
321	7			321	7
322	9	--INTERPOLACIÓN -->		322	9
323	11			323	11
324	NaN			324	11
325	NaN			325	11
326	NaN			326	11

Figura 28. Ejemplo valores NaN al final

En la Figura 28 se puede observar como existen 3 valores *NaN* al final y no existe un valor siguiente. Por lo que se elige el valor anterior válido para sustituir los valores *NaN*, que en este caso es el 11.

A continuación, se va a explicar cómo se ha llevado a cabo esta interpolación en los datos de este estudio.

Se va a explicar cómo se ha implementado en Python esta interpolación para una columna.

Para la interpolación implementada, se van a recorrer todas las filas de la columna. Se debe comprobar si la primera iteración del bucle es un *NaN*. Si lo es, se trata de un caso especial de los comentados. Se debe buscar el siguiente valor válido, ya que se va a sustituir por el/los valor/es *NaN*. Para encontrar el siguiente valor válido se va a utilizar la función *first_valid_index()*, la cual devuelve la posición del siguiente valor no *NaN* de la columna. A esta posición siguiente se le va a nombrar "s_i", y al valor de esa posición, que es el valor siguiente se le va a nombrar "s". Y teniendo la posición, es fácil obtener el valor utilizando "*df.iloc[filas, columnas]*".

La nomenclatura de la posición de un valor válido anterior y el valor anterior en sí, es muy parecida. En vez de la "s" de "siguiente" se utiliza una "a" de "anterior". Por lo que la posición de un valor válido anterior es "a_i" y el valor válido anterior es "a".

Cuando se obtiene en la primera posición un *NaN* (o varios *NaN* en las primeras posiciones), no existe "a_i", ni "a". Pero el objetivo es repetir el primero no nulo desde "a_i" que es cero, hasta "s_i", que ha sido calculado con: *first_valid_index()*.

Fuera del caso especial, lo que se va a hacer es iterar todas las filas y en cada iteración comprobar si ese valor es *NaN*, utilizando la función "*np.isnan*". En caso de encontrar un nulo, la variable "a_i" toma el valor de la posición anterior a la iteración, es decir "a_i = i – 1". Ese valor anterior se puede obtener con la posición de este, con *iloc*. Y la posición del siguiente no nulo se obtiene al igual que antes, con *first_valid_index()* y su valor también con *iloc*. Y el número de nulos se obtiene haciendo la diferencia de las posiciones del siguiente y el anterior: "numero_nulos= s_i – a_i -1"

Una vez se tienen todas las variables con valor, se comprueba que la posición del valor siguiente existe. Ya que si "s_i = *None*" significa que no hay un valor siguiente válido; sino que hay todos valores *NaN* hasta el final de la columna. Este es el otro caso especial que se ha nombrado. Lo que se debe hacer en este caso, es utilizar el valor anterior válido para sustituirlo en los *NaN* del final haciendo: "*df.iloc[i : numero_filas] = a*"; siendo "numero_filas" el total de filas de la columna. Por lo que se está introduciendo el valor de "a" en todas las posiciones desde la posición actual, hasta el final de la columna.

Y por otra parte, si "s_i" no es "*None*" es el caso habitual de interpolación, en el que se aplica la formula. Y en este caso se implementa un bucle ya que si "numero_nulos" es mayor que 1, significa que hay varios valores *NaN* y hay que interpolar varios valores *NaN* consecutivos. Y la "i" de esta formula: Valor interpolado = a + i x va aumentando hasta "numero_nulos" (en el código implementado), que es "n" (en la formula).

Esta interpolación ha sido explicada para una columna, pero hay que interpolar 5 columnas ('power', 'poa', 'poa2', 'Tcell' y 'Tcell2'), como ya se había mencionado. Por lo que todo lo anterior se aplica a las 5 columnas, diseñando un bucle for de 5 iteraciones.

Los *dataframes* con valores interpolados se van a nombrar "dfX_i"

4.2.9 No interpolación

Esta es la otra opción que se debe tomar, si no se desea implementar la interpolación. Esta es más sencilla ya que consiste en eliminar datos. Cualquier fila que

contenga un dato inválido o *NaN*, debe ser eliminada. Esta opción se ha descrito como "No interpolación" pero realmente se podría llamar: "Limpieza de datos inválidos". Pero se ha nombrado así para aclarar, que existen dos soluciones distintas. Hasta ahora la fase de preprocesado era un proceso detrás de otro, pero ahora surgen dos ramas o alternativas distintas. Se implementan las dos y posteriormente se analizará que opción es la más válida para este estudio.

Esta fase de limpieza de datos se puede implementar con la función *dropna()*.

Se puede calcular el número de filas antes de aplicar la limpieza de datos y después y comprobar que se han eliminado las 152278 filas con valores *NaN*. Los *dataframes* sin valores *NaN* se van a nombrar "dfX_sin_nulos"

4.2.10 Media de datos de sensores

Las columnas: 'poa y 'poa2' contienen las medidas de 2 sensores de radiación situados en el plano del panel. En un mismo campo solar la radiación captada de estos dos sensores debe ser muy parecida. Realmente con una medida sería suficiente, pero una de las columnas está como complementaría. Lo mismo sucede con las columnas: 'Tcell' y 'Tcell2', las cuales contienen medidas de 2 sensores de temperatura colocados en el plano del panel solar. En un mismo campo solar, estas dos temperaturas también deben ser muy similares entre ellas.

Tras haber finalizado las fase de interpolación y la limpieza de datos, se va a proceder a disminuir el número de columnas. Se pretende realizar la media de las columnas de radiación y temperatura. En vez de tener 4 columnas de medidas de sensores, se van a tener solo dos columnas: 'media_poa' y 'media_temp'.

Este proceso se va a aplicar tanto en los datos interpolados como en los no interpolados. Es decir sobre los *dataframes* "dfX_i" (interpolados) y sobre los *dataframes* "dfX_sin_nulos" (no interpolados). Se habla de "*dataframes*" en plural ya que se aplica a todos los campos solares, siendo X el número del campo solar.

A esta función se ha definido como "media_poa_temp" y consiste en seleccionar las columnas tercera y cuarta ('poa' y 'poa2'), hacer la media utilizando la función

mean(), y añadir una nueva columna ('media_poa') que sea el resultado de la media. Lo mismo para la temperatura. Se selecciona las columnas quinta y sexta ('Tcell' y 'Tcell2') y se hace la media de estas dos columnas y se añade el resultado en una nueva columna ('media_temp').

Y por último se eliminan las 4 columnas: ('poa', 'poa2', Tcell' y 'Tcell2') que son las columnas con las que hemos calculado la media.

El nuevo *dataframe* que contiene la media de los sensores de radiación y temperatura se ha definido como "dfX_i_m" para los datos que han sido interpolados y "dfX_ni" para los datos que no han sido interpolados (para los que han sido eliminados). Y en ambos casos el número de columnas de los *dataframes* es el mismo. Se trata de 4 columnas: 'Timestamp', 'power', 'media_poa' y 'media_temp'.

4.2.11 Exportar datos

Para terminar con la fase de preprocesado, los datos que han sido modificados van a ser almacenados como archivos csv, para que no sea necesario ejecutar esta fase cada vez que se quiere ejecutar una fase de las que se van a estudiar a continuación.

Esto se debe a que la fase de preprocesado tiene un largo tiempo de ejecución, debido a la interpolación, ya que tiene que hacer muchos cálculos y pasa por todas las filas y por todas las columnas menos la de la fecha.

Si se almacena cada campo solar como un archivo csv, luego tan solo hay que importar esos datos y no volver a ejecutar la fase de preprocesado de nuevo, la cual se demora.

Así pues, se va a utilizar la función *to_csv()* para almacenar los *dataframes* como archivos csv. Como existen 4 campos solares y se han implementado dos alternativas con ellos (interpolación y no interpolación), el resultado son 8 *dataframes* (2 *dataframes* con soluciones distintas por cada campo solar). Es decir se almacenan los 4 *dataframes* de interpolación: "df1_i_m", "df2_i_m", "df3_i_m" y "df4_i_m" y los 4 *dataframes* de no haber interpolado: "df1_ni", "df2_ni", "df3_ni" y "df4_ni".

Y antes de almacenar estos *dataframes*, como comprobación de si se han realizado bien las fases de interpolación y no interpolación, se va a comprobar si existe algún valor inválido en los *dataframes* ejecutando "*dfX_ni.isnull().values.any()*" y "*dfX_i_m.isnull().values.any()*". Si devuelve *False* significa que no hay ningún nulo y si devuelve *True* significa que hay algún nulo.

Una vez se ha comprobado que no hay ningún nulo porque nos devuelve False, se almacenan los *dataframes* como archivos csv con los nombres: "dfX_i.csv" para interpolados y "dfX_ni.csv" para los no interpolados.

En la Figura 29 se puede observar como ha sido la evolución del número de filas y se aprecia cómo estas se han ido reduciendo hasta obtener los datos que finalmente han sido exportados.

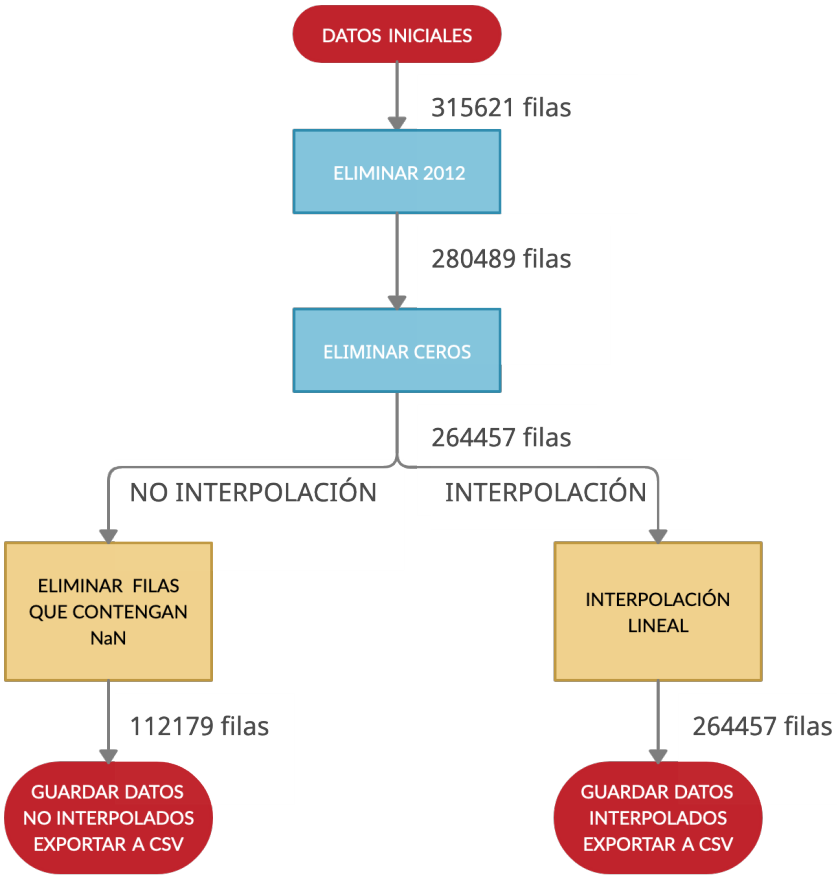

Figura 29. Evolución del número de filas

4.3 Visualización de los datos

En esta fase del estudio de la degradación se va realizar un análisis de los datos ya preprocesados, tanto de los datos interpolados como los no interpolados. Este análisis trata de estudiar los datos de una manera más estadística para saber cómo los datos están distribuidos, cómo se relacionan las diferentes variables entre ellas, cuál es el rango de valores de cada variable, cuál es la media y la mediana para cada variable, etc.

Para ello, lo primero que se tiene que hacer es importar los archivos csvs como *dataframes*. Los *dataframes* de los datos interpolados están definidos como "dfX_i" y los *dataframes* de los datos no interpolados están definidos como "dfX_ni". Estos *dataframes* tienen las siguientes columnas: 'Timestamp', 'power', 'media_poa' y 'media_temp'. Y por cuestión de facilidad en la nomenclatura, se van a renombrar estas dos últimas columnas. La columna 'media_poa' se cambia por simplemente 'poa' y la columna 'media_temp' se cambia por simplemente 'Tcell'.

A continuación se va a utilizar la función "describe()" sobre el campo solar 1 tanto para los datos interpolados como para los no interpolados. Se va a realizar solo sobre el campo solar 1 porque los datos van a ser similares para los demás campos. El objetivo es comprender los datos y ver las diferencias que existen entre los datos que hemos conseguido tras el preprocesado con interpolación y el preprocesado sin interpolación.

La función describe() muestra un resumen con detalles estadísticos del *dataframe* para cada columna de valores numéricos. En esta estadística está incluido: "*count*" que es el número de valores de esa columna, "mean" que es la media de esa columna, "*std*" que es la deviación típica de esa columna, "min" que es el mínimo de esa columna, "*max*" que es el máximo de esa columna, y ("25%", "50%" y "75%") que son los cuartiles.

Así pues, ejecutando "df1_i.describe()" se obtiene el resumen estadístico del campo solar 1 tras un preprocesado con interpolación que se puede observar en la Tabla 1.

	power	*poa*	*Tcell*
Valores	*264457*	*264457*	*264457*
Media	*40710.92*	*222.38*	*20.80*
Desviación	*51164.44*	*305.88*	*13.47*
Mínimo	*30.51*	*-0.01*	*-5.68*
25 %	*627.34*	*0.00*	*10.43*
50 %	*10141.20*	*32.13*	*17.93*
75 %	*78558.65*	*397.34*	*29.23*
Máximo	*174929.08*	*999.99*	*65.31*

Tabla 1. Estadísticas df1 interpolado

Y se realiza lo mismo para los datos no interpolados. Ejecutando "df1_ni.describe()" se obtiene el resumen estadístico del campo solar 1 tras un preprocesado sin interpolación que se puede observar en la Tabla 2.

	power	*poa*	*Tcell*
Valores	*112179*	*112179*	*112179*
Media	*70589.43*	*413.08*	*28.92*
Desviación	*53239.20*	*312.85*	*12.66*
Mínimo	*30.51*	*3.02*	*-3.59*
25 %	*18490.76*	*114.15*	*18.91*
50 %	*64763.21*	*362.53*	*28.21*
75 %	*121005.54*	*691.51*	*38.64*
Máximo	*172690.94*	*999.99*	*63.79*

Tabla 2. Estadísticas df1 no interpolado

Como era de esperar, los datos no interpolados tienen menos datos que los datos interpolados. El número de columnas es el mismo para ambos, pero el número de filas no. Ya que los datos no interpolados se han borrado todas las filas con datos de 2012, las filas con potencia igual a cero y las filas que tuviesen algún valor inválido (*NaN*). Sin embargo los datos interpolados, se han borrado todas las filas con datos de 2012 y las filas con potencia igual a cero; pero las filas que tienen algún valor inválido han sido predichas linealmente o interpoladas. Por lo que tiene sentido que los datos interpolados tengan mayor número de filas. Esto se puede comprobar en la Tabla 3, que es una comparativa del número de valores que tiene el campo solar 1 para datos interpolados y para no interpolados.

		power	*poa*	*Tcell*
Valores	*Interpolado*	*264457*	*264457*	*264457*
Valores	*No interpolados*	*112179*	*112179*	*112179*

Tabla 3. Comparación valores df1

Existe una gran diferencia entre la media de los datos interpolados y los datos no interpolados. Como se puede observar en la Tabla 4, la media de los datos interpolados es mucho menor que la media de los datos no interpolados en las 3 variables de interés.

		power	*poa*	*Tcell*
Media	*Interpolado*	*40710.92*	*222.38*	*20.80*
Media	*No interpolados*	*70589.43*	*413.08*	*28.92*

Tabla 4. Comparación media df1

4.3.1 Distribución de los datos

Así que para entender mejor este efecto se va a representar un histograma para ver la distribución de los datos. Para hacer un histograma, se debe contar las veces que aparece cada valor en cada intervalo en el conjunto de valores

Los histogramas sirven para observar los intervalos de valores que están más presentes en los datos, es decir para ver si existen datos más frecuentes que otros. Los histogramas entonces, son gráficos que muestran la frecuencia de los datos mediante una distribución de estos. Se trata de representar gráficamente una variable en forma de barras. Y la superficie de cada barra es proporcional a la frecuencia de los valores. Las barras o rectángulos todos tienen el mismo ancho ya que los intervalos son definidos y lo que varía es la altura que hace referencia a la cantidad de datos dentro de ese intervalo.

Los histogramas se pueden hacer con "*pyplot*" haciendo uso de la función "*hist*", a la que se le pasa: la información que queremos analizar, el número de intervalos en el que se desea clasificar los valores (*bins*) (es opcional, ya que por defecto son 10 intervalos), y otra información no tan relevante.

También se puede pasar una lista que contenga los límites de los intervalos, en vez de pasar a la función el número de intervalos, al parámetro "*bins*". En este estudio para cada variable, se ha pasado una lista diferente ya que los rangos de valores de cada variable son diferentes.

La potencia tiene un rango (también conocido como recorrido) de valores desde 30.51 W hasta 172690.94 W (para los datos no interpolados): la potencia puede tomar aproximadamente 172660 valores.

Como cada variable tiene un rango diferente, tiene sentido que los intervalos definidos para ellas no sean los mismos. Por lo que se van a definir 3 listas para crear los intervalos de los histogramas; una lista por cada variable (potencia, radiación y temperatura). Para crear estos intervalos para representar los histogramas es necesario basarnos en el recorrido de la variable. Así que, se ha utilizado el máximo y el mínimo de cada variable para crear estas listas que van a crear los intervalos.

Por lo que para la temperatura, se ha implementado una lista desde 0 hasta 175 mil con saltos de 10 mil unidades (o lo que es lo mismo: paso = 10 mil) y es la que se ha pasado como parámetro "*bins*". Por lo que se van a contar el número de valores que hay entre 0 y 10 mil y se va a representar, los valores que hay entre 10 mil y 20 mil y se va a representar, y así hasta 175 mil.

Si una barra del histograma alcanza valores verticales muy elevados significa que en los datos analizados, existen muchos valores entre los rangos del rectángulo que forman el intervalo. Sin embargo si estos rectángulos son más bajos significa que existen pocos valores en ese intervalo.

Se puede apreciar en la Figura 30, la cual representa la distribución de los datos no interpolados, que los valores de potencia están distribuidos de una manera muy uniforme. Esto es una buena señal ya que si en los histogramas existe una gran diferencia de valores en los intervalos, puede ser que estos valores se traten de valores de medidas erróneas (*outliers*). También se puede destacar, que esta distribución tiene gran cantidad de datos (alrededor de 19 mil valores) cuya potencia se encuentra entre 0 y 10 mil vatios (10 kW).

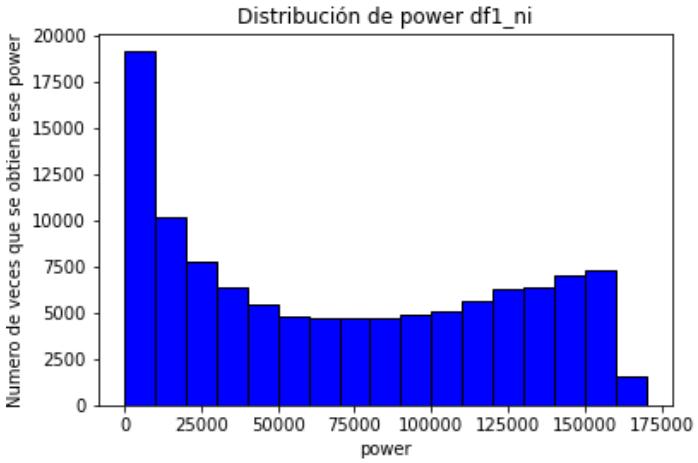

Figura 30. Histograma de potencia df1 no interpolado

Sin embargo, si se observa la distribución de la potencia en la Figura 31 que corresponde con los datos interpolados, es de notar que existe una enorme diferencia entre el rectángulo del intervalo inicial y el resto. En la potencia de los datos interpolados,

hay 135 mil valores que tienen una potencia entre 0 y 10 kW. El número total de valores interpolados es de casi 265 mil, por lo que se puede afirmar que más de la mitad de los datos de potencia tienen valores muy pequeños de potencia (entre 0 y 10 kW). Lo cual, es un indicador de que la interpolación lineal que se ha realizado en la fase de preprocesado no es muy fiable. Quizás al interpolar había muchos valores cero. Y cuando se interpolan varios valores *NaN* entre dos valores válidos que son cero, los datos que se están añadiendo son también 0s. Este efecto hace que la distribución sea menos constante y más radical.

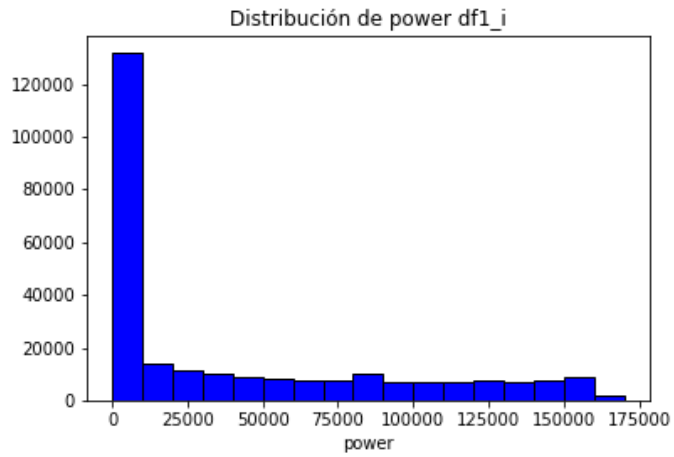

Figura 31. Histograma de potencia df1 interpolado

En la Figura 31 que corresponde a los datos interpolados se puede observar que se trata de un histograma sesgado. El histograma es anormalmente alto al principio de la distribución. Esta distribución es asimétrica ya que la curva termina bruscamente en uno de los lados y hacia el otro lado la curva es suave.

Cuando los datos son muy homogéneos, como ocurre en la Figura 30 la media aporta un valor representativo de la realidad y por lo tanto en la Tabla 4 las medias de los datos no interpolados si que representan la realidad, pero cuando los datos son muy heterogéneos, como ocurre en la Figura 31 la media no aporta un valor representativo de la realidad, ya que los datos no están distribuidos de manera equilibrada. Por lo tanto, si se observa la Tabla 4 las medias de los datos interpolados no aportan una perspectiva de cómo son los datos realmente. Este efecto se ve reflejado en la Tabla 4, donde se observa que las medias para los datos interpolados son mucho más bajas que para los

no interpolados. Este diferencia tan grande entre las medias se entiende mejor cuando se comparan la Figura 30 y la Figura 31 y se ve la gran cantidad de valores pequeños de potencia (la barra primera del histograma es muy alta al principio de la distribución en la Figura 31) que existen en los datos interpolados, lo cual hace que disminuya la media de todo el conjunto.

A continuación se va a representar un histograma para la radiación. Para así también realizar una comparación de distribuciones entre los datos interpolados y los no interpolados y poder entender por qué existe tanta diferencia entre las medias de los datos interpolados y los no interpolados.

Como el mínimo de la radiación es aproximadamente 3 y el máximo 1000, la lista de radiación se ha definido como: "*lista= list(range(0,1000,50))*".

Y en la Figura 32 que representa el histograma del campo solar 1 de los datos no interpolados, se puede observar como al igual que ocurría en la potencia; este histograma tiene una distribución muy uniforme. Está un poco sesgada a la derecha (un poco más que la de potencia). Los valores de radiación más frecuentes están entre valores de 0 y 50, habiendo unos 15 mil valores en ese intervalo. Y no se puede destacar solo un intervalo menos frecuentado ya que casi todos los intervalos rondan el mismo número de valores, por lo que los datos sin interpolación están bastante equilibrados, aunque tenga intervalos menos frecuentados como los dos intervalos del medio que abarcan el rango de 450 a 550 .

El sesgo hacia a la derecha (existe una primera barra con gran cantidad de valores en la distribución) que se ha observado en la distribución de la potencia (en la Figura 31) y en la distribución de la radiación (en la Figura 32) es razonable que exista, ya que es normal que por la noche la radiación sea muy pequeña y que por lo tanto la potencia que se obtenga también sea muy baja. Esta representación de datos es causa de la noche, pero además se pueden añadir otros efectos que no tienen por que ocurrir por la noche, como el efecto de ensuciamiento o que salga un día muy nublado y los valores de radiación sean muy bajos.

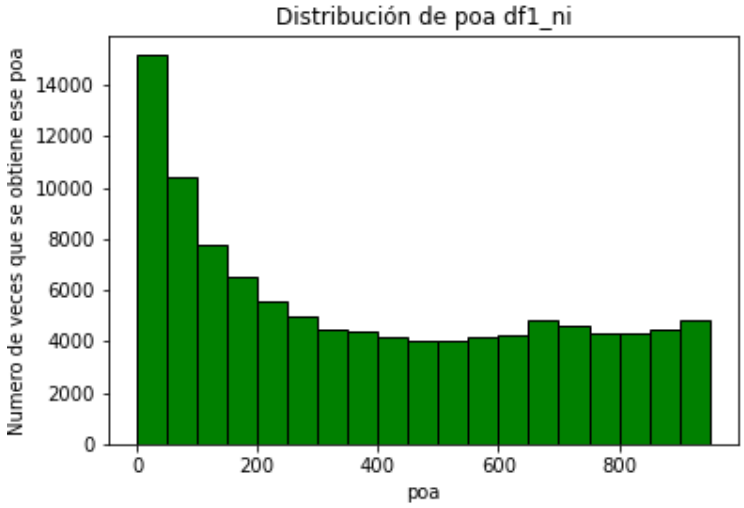

Figura 32. Histograma de radiación df1 no interpolado

Seguidamente se realiza la representación gráfica de la distribución de los datos interpolados, en la que se aprecia el mismo comportamiento que en la temperatura.

Este histograma indica que existen casi 140 mil valores entre el intervalo de radiación: 0 y 50 . Sin embargo había unos 15 mil valores en ese intervalo para los datos no interpolados.

Si se compara el primer intervalo con el resto existe una diferencia de valores grandísima. Para el primer intervalo se encuentran 140 mil valores y el resto de los intervalos, la mayoría no llega ni a los 10 mil valores. Por tanto la distribución de la Figura 33 no se trata de una distribución uniforme de la cual se puedan obtener resultados fiables.

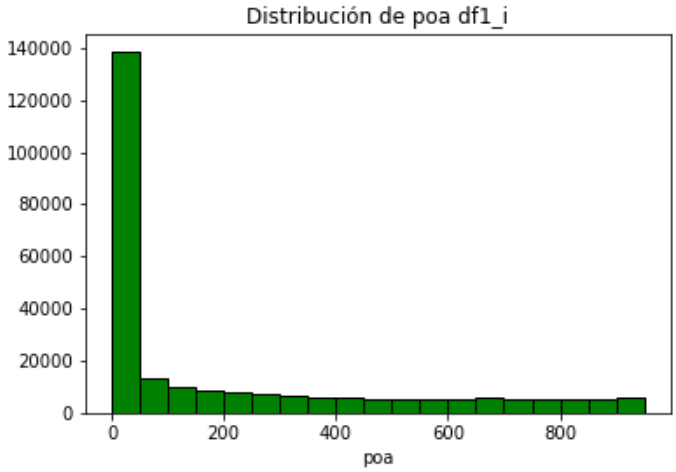

Figura 33. Histograma de radiación df1 interpolado

Por definición, las distribuciones sesgadas hacia la derecha son las que tienen la media mayor que en la mediana. Y esto es lo que ocurre en los datos de la Tabla 5. Cuanto mayor sea la diferencia entre media y mediana, el sesgo va a ser mayor. Por eso para los datos interpolados, los histogramas de potencia y radiación están más sesgados que el histograma de temperatura. Porque la diferencia entre la media y la mediana de temperatura no es tan grande.

Cabe destacar que las medianas de la Tabla 5 y la Tabla 6 no iban incluidos dentro del resumen estadístico que proporciona la función "describe", sino que se ha tenido que utilizar la función "median" para calcularla.

	power	*poa*	*Tcell*
Media df1_i	*40710.92*	*222.38*	*20.80*
Mediana df1_i	*10141.32*	*32.13*	*17.93*

Tabla 5. Media vs mediana datos interpolados

Sin embargo en los datos no interpolados la media y la mediana son muy similares como se puede ver en la Tabla 6, aunque la media sea superior a la mediana en las 3 variables: potencia, radiación y temperatura.

	power	*poa*	*Tcell*
Media df1_ni	*70589.43*	*413.08*	*28.92*
Mediana df1_ni	*64763.21*	*362.53*	*28.21*

Tabla 6. Media vs mediana datos no interpolados

Y por último se representan los histogramas para la variable que faltaba, para la temperatura. En este caso, el recorrido de la temperatura para los datos no interpolados es de -3.59ºC a 63.79ºC .Por lo que se ha creado la lista de intervalos entre -10 y 70 con un paso de 5 unidades, quedando la lista tal que así: "lista= list(range(-10,70,5))".

Como se puede ver en la Figura 34, esta distribución no es asimétrica como las anteriores, es más bien simétrica. Las anteriores tenían sesgo hacia la derecha y esta distribución se asemeja a una normal. En una distribución normal, la media y la mediana son muy similares, y en este caso lo son. Ya que la media es 28.92ºC y la mediana 28.21ºC.

La mayoría de valores se encuentran entre los 20 y 30ºC. Y los valores que son menos frecuentes, son las temperaturas negativas y las temperaturas mayores de los 60ºC.

Para la potencia y radiación, que tenían una distribución heterogénea, no es muy recomendable fiarse de la media, sino que es mejor usar la mediana ya que esta va a informar mejor el punto central de la distribución porque está menos afectada por la presencia de sesgos. Pero para la temperatura, la cual tiene una distribución normal si que es recomendable fiarse de la mediana.

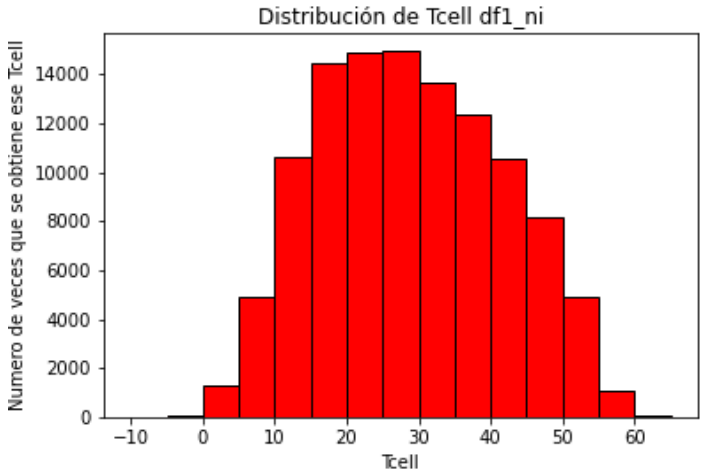

Figura 34. Histograma de temperatura df1 no interpolado

El histograma de temperatura de los datos interpolados que se puede observar en la Figura 35, no tiene una distribución muy distinta al histograma de temperatura de los datos no interpolados de la Figura 34. En cambio, con el resto de variables, potencia y radiación sí que había un cambio enorme en la distribución. En este caso, la normal de temperatura de los datos no interpolados se ha sesgado un poco en los datos interpolados. Pero este sesgo no es como el de la potencia y la radiación donde tan solo había un intervalo muy alto al principio, sino que se mantiene la distribución que comienza con intervalos poco poblados, luego aumentan los datos en los intervalos intermedios y finalmente los intervalos finales vuelven a estar poco poblados. Es decir, los intervalos de los extremos están poco poblados.

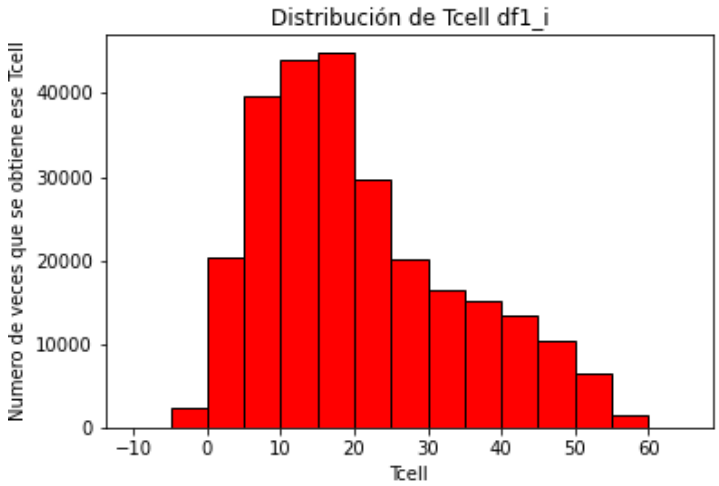

Figura 35. Histograma de temperatura df1 interpolado

Este menor cambio del histograma de temperatura entre datos no interpolados e interpolados se podría ver reflejado en la diferencia de las medias. Se puede utilizar la Tabla 4 para realizar la Tabla 7. En esta última, se han utilizado las medias de ambos datos y el recorrido de la variable para calcular cuál es la diferencia entre las medias.

		power	*poa*	*Tcell*
Media	*Interpolado*	*40710.92*	*222.38*	*20.80*
Media	*No interpolados*	*70589.43*	*413.08*	*28.92*

Tabla 7. Diferencia (%) de media interpolada y media no interpolada.

Los resultados de lo formulado en la Tabla 7, se encuentran en la Tabla 8 . Esta diferencia se ha multiplicado por 100 para obtener el resultado en forma porcentual.

En esta Tabla 8 se puede apreciar que las variables con mayor diferencia entre las medias son la potencia y la radiación con un 17% y 19%. Sin embargo la media no cambia tanto en la temperatura, que solo cambia un 12% entre datos no interpolados e interpolados. Por eso, se entiende que existan mayores diferencias en la distribución de los histogramas de no interpolados y de interpolados para las variables de temperatura y radiación, y que los histogramas de temperatura sean más parecidos entre sí.

		power	*poa*	*Tcell*
Media	*Interpolado*	*40710.92*	*222.38*	*20.80*
Media	*No interpolados*	*70589.43*	*413.08*	*28.92*
DIFERENCIA		*17.3 %*	*19.12 %*	*12.04%*

Tabla 8. Resultados diferencia (%) de media interpolada y media no interpolada.

Los datos de la Tabla 8 para la potencia son representados gráficamente en la Figura 36, pero no como una media global de todo el conjunto de datos, sino la media por años. Estas gráficas se han representado para todos los campos solares y para las tres variables y en todas ellas se aprecia esa amplia diferencia entre las medias. Por lo que no ha sido necesario incluir todas estas gráficas en este estudio.

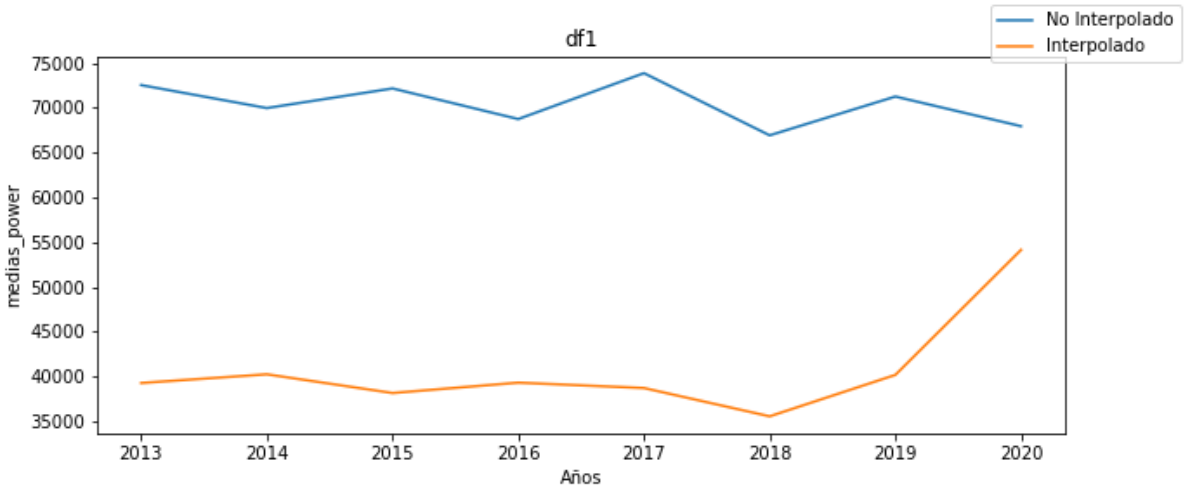

Figura 36. Comparación medias datos interpolados vs no interpolados

Para calcular la diferencia entre la media de interpolados y la media de no interpolados de la Tabla 8 ha sido necesario utilizar los datos de la Tabla 9 y Tabla 10 para hacer la diferencia entre máximo y mínimo y saber el recorrido de las variables.

		power	*poa*	*Tcell*
Mínimo	*Interpolado*	*30.51*	*-0.01*	*-5.68*
Mínimo	*No interpolados*	*30.51*	*3.02*	*-3.59*

Tabla 9. Comparación mínimo df1

		power	*poa*	*Tcell*
Máximo	*Interpolado*	*174929.08*	*999.99*	*65.31*
Máximo	*No interpolados*	*172690.94*	*999.99*	*63.79*

Tabla 10. Comparación máximo df1

La desviación estándar es la media de la distancia de todas las muestras a la media de todos los datos. Si se obtiene una desviación muy alta, significa que los datos están muy dispersos respecto a su valor medio. En la Tabla 11 se observan las desviaciones estándar que se han obtenido en las estadísticas de la función "describe()".

		power	*poa*	*Tcell*
Desviación	*Interpolado*	*51164.44*	*305.88*	*13.47*
Desviación	*No interpolados*	*53239.20*	*312.85*	*12.66*

Tabla 11. Comparación desviación df1

En el resumen de estadísticas de "describe" también se incluían los cuartiles. Los cuartiles son los valores que delimitan la división de los datos en cuatro grupos que contienen el mismo número de valores. El 100% de los datos se divide en cuatro partes iguales: 25%, 50%, 75% y 100%. El primer cuartil indica que el 25% de los valores son menores o igual a este valor. El segundo cuartil corresponde con la mediana y divide en dos partes iguales la distribución. De forma que el 50% de los valores son menores o igual al valor del segundo cuartil. El tercer cuartil indica que el 75% de los valores son menores o igual a este valor. El cuartil uno es conocido como Q1, el cuartil dos como Q2 o mediana, el cuartil 3 como Q3. Los datos correspondientes a los cuartiles se han detallado en la Tabla 12, en la Tabla 13 y en la Tabla 14, haciendo una comparativa de los cuartiles para los datos interpolados y de los no interpolados. La Tabla 12 corresponde con el cuartil uno (Q1), donde se encuentran el 25% de los datos hasta el valor del cuartil.

		power	*poa*	*Tcell*
25 %	*Interpolado*	*627.34*	*0.00*	*10.43*
25 %	*No interpolados*	*18490.76*	*114.15*	*18.91*

Tabla 12. Comparación 25 % df1

En la Tabla 13, que corresponde con el cuartil dos (Q2), se puede apreciar como estos valores corresponden con la mediana si se observa la Tabla 5 y la Tabla 6. Hasta el valor del cuartil dos se encuentran el 50% de los datos.

		power	*poa*	*Tcell*
50 %	*Interpolado*	*10141.20*	*32.13*	*17.93*
50 %	*No interpolados*	*64763.21*	*362.53*	*28.21*

Tabla 13. Comparación 50% df1

La Tabla 14 corresponde con el cuartil tres (Q3), donde se encuentran el 75% de los datos hasta el valor del cuartil.

		power	*poa*	*Tcell*
75 %	*Interpolado*	*78558.65*	*397.34*	*29.23*
75 %	*No interpolados*	*121005.54*	*691.51*	*38.64*

Tabla 14. Comparación 75% df1

La explicación de los cuartiles se puede entender mejor, estudiando los diagramas de cajas o "*Boxplot*". Se trata de unas gráficas que describen la dispersión y la simetría de la distribución de los datos. Los diagramas de cajas representan gráficamente los 3 cuartiles (Q1,Q2 y Q3), el mínimo y el máximo.

Está compuesto por un rectángulo o caja, que representa los valores de los 3 cuartiles. Y además tiene dos bigotes uno a la izquierda de la caja, que representa el mínimo, y otro a la derecha de la caja, que representa el máximo. Esto se puede ver representado en la Figura 37, en la que el mínimo sería la línea vertical que dice: "valor más bajo" y el máximo sería la línea vertical que dice: "valor más alto".

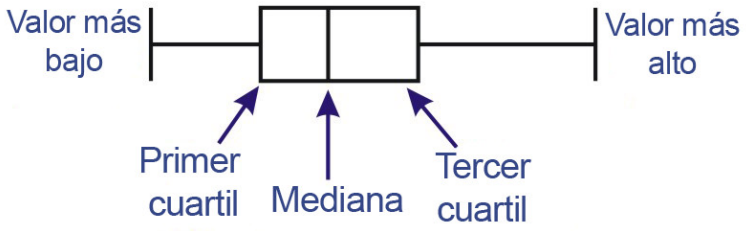

Figura 37. Esquema del diagrama de caja [25]

Esta representación se divide en 4 segmentos. Por lo que los datos van a estar divididos en 4 grupos de 25% cada uno. El 25% de los datos se encuentra el mínimo y Q1, otro 25% de los datos se encuentra entre Q1 y Q2, otro 25% de los datos se encuentra entre Q2 y Q3, y el último 25% de los datos se encuentra entre Q3 y el máximo.

Los bigotes de la caja pueden servir para identificar valores atípicos, los cuales están más alejados de los valores máximo y mínimo.

En este estudio para representar los diagramas de cajas en Python se ha utilizado la librería: "*Seaborn*".

Lo habitual es que la distribución de los datos no sea exactamente simétrica, y por lo tanto la distancia entre los cuartiles no sea la misma. Sin embargo unos datos bien distribuidos implica que la distancia entre las 4 barras verticales sea la misma (Eso es el caso ideal). Y como se puede observar en la Figura 38, el diagrama de caja de la potencia de los datos no interpolados están distribuidos de forma homogénea.

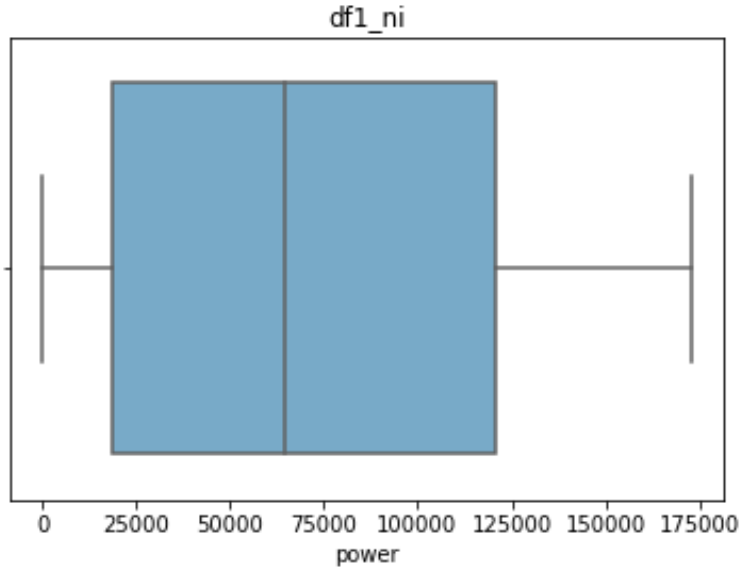

Figura 38. Diagrama de caja potencia datos no interpolados

La distribución de los datos que se ve en la Figura 39 es también la de la potencia pero de los datos interpolados. Sin embargo este diagrama de caja no está bien compensado ya que el primer 25% y el segundo 25% se encuentra en valores muy bajos. Se trata de una distribución de los datos muy heterogénea, la cual índica que existen muchos valores de potencia bajos.

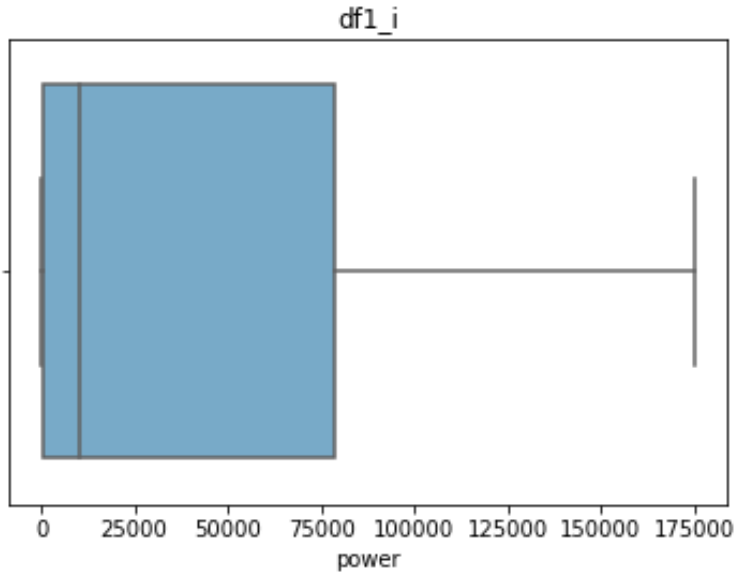

Figura 39. Diagrama de caja potencia datos interpolados

Lo mismo ocurre para la radiación. El diagrama de caja de los datos interpolados es el esperado, con una distribución de datos homogénea; como se ve en la Figura 40 y la de los datos interpolados, que es la de la Figura 41, todo lo contrario. Por lo que los datos interpolados no aparentan ser datos que puedan ser útiles para este estudio debido a su baja fiabilidad.

Por lo tanto, para el aprendizaje automático se va a continuar con los datos no interpolados, ya que son los más fiables.

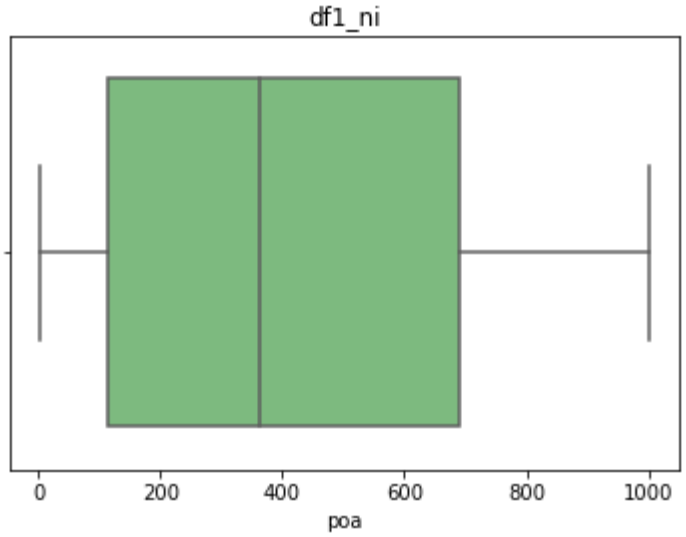

Figura 40. Diagrama de caja radiación datos no interpolados

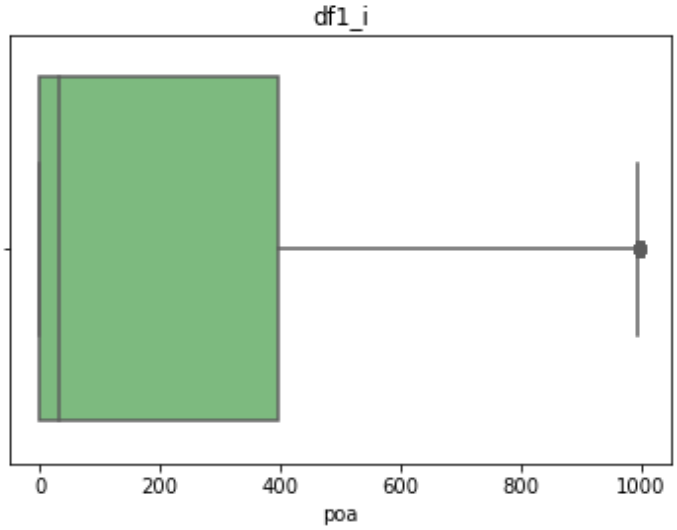

Figura 41. Diagrama de caja radiación datos interpolados

Para la temperatura ambos diagramas de cajas son similares. Ambos tienen una distribución homogénea, sobre todo la de los datos no interpolados de la Figura 42. Cuando una variable se ve tan bien distribuida, si se representase en un histograma se apreciaría que esta variable sigue una distribución normal, como ocurre en la Figura 35.

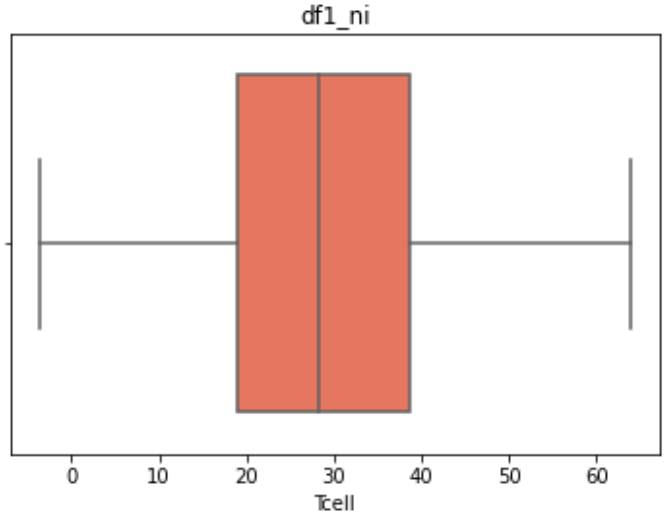

Figura 42. Diagrama de caja temperatura datos no interpolados

Por otro lado, aunque la distribución de los datos interpolados de la Figura 43 sea homogénea, existen unos valores que están por encima del bigote del máximo. Esto son valores atípicos que se encuentran en los datos y que solo introducen ruido al estudio en cuestión. Por lo que este diagrama de caja muestra que no todos los datos interpolados son fiables ya que existen *outliers* en ellos.

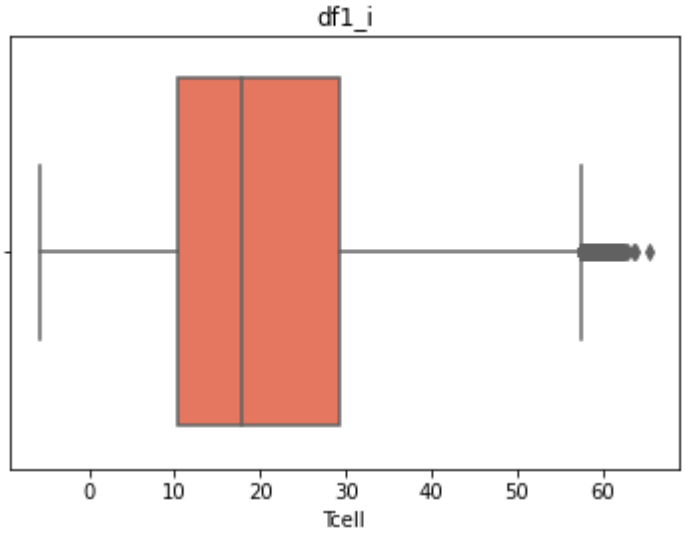

Figura 43. Diagrama de caja temperatura datos interpolados

Todo esta fase de estadística se ha estado hablando sobre que los datos interpolados están sesgados y se ha comprobado numéricamente que esto es debido a que la media es mucho mayor a la mediana. Ahora se va a apreciar visualmente este efecto, ya que se va a representar gráficamente la media y la mediana para todos los años. Así que, se puede observar esta gran diferencia entre la media y la mediana de la potencia para los datos interpolados del campo solar 1 en la Figura 44.

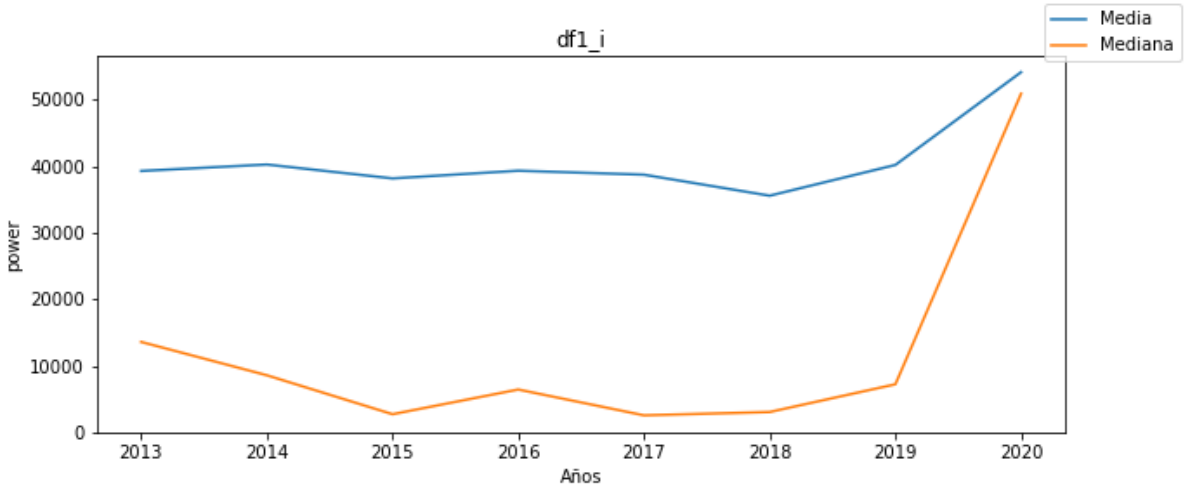

Figura 44. Potencia media vs mediana datos interpolados

En cambio para los datos no interpolados, se ha realizado la misma comparación entre media y mediana, y se puede observar en la Figura 45 que ambas siguen la misma tendencia con el paso de los años, ya que se tratan de datos distribuidos uniformemente.

Figura 45. Potencia media vs mediana datos no interpolados

La Figura 44 y la Figura 45 son dos ejemplos de gráficas para entender mejor el por qué de elegir los datos no interpolados para este estudio, en vez de los interpolados. Ya que esto no ocurre solamente para el campo solar uno y con la potencia. Estas gráficas se han realizado para los 4 campos solares y para las 3 variables, obteniendo las mismas conclusiones en todas ellas; pero estas no van a ser añadidas a la memoria del estudio por no ser repetitivos.

En la Figura 46 parece que la línea de la media de Tcell de los datos interpolados es más estable que la línea de la media de Tcell con datos no interpolados; ya que tiene menos picos.

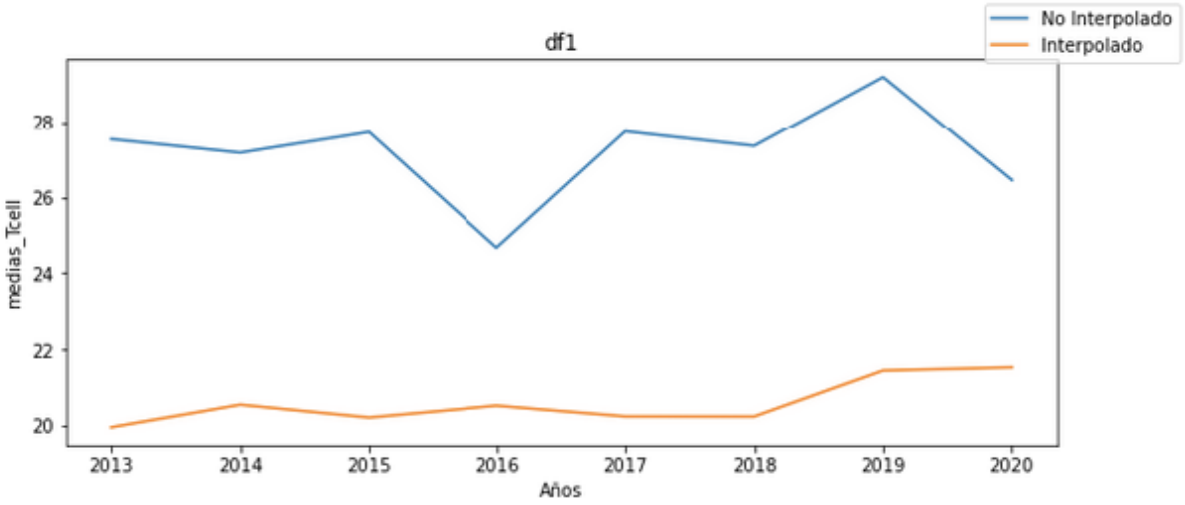

Figura 46. Temperatura de datos interpolados y no interpolados

Esta gráfica puede hacer que en este estudio se seleccionen los datos interpolados ya que parecen más fiables. Sin embargo, la diferencia entre las dos medias es grande (casi 8ºC en un recorrido de temperatura de 70ºC). Por lo tanto, la media interpolada es más estable porque se añaden gran cantidad de valores pequeños a los datos al interpolar en todos los años, y esto hace que los picos que existen sean moderados.

Una vez que se ha comprobado mediante el análisis de datos interpolados y no interpolados, que los datos más fiables son los datos no interpolados. A pesar de que los datos no interpolados era la opción simple de implementar, ya que los datos interpolados han llevado su tiempo; tanto tiempo de implementación del propio algoritmo como tiempo de ejecución. Pero gracias al estudio de los datos se ha podido concluir que conviene utilizar los datos no interpolados. Por lo que en el estudio se van a utilizar los datos no interpolados desde aquí en adelante. Así que en las siguientes fases aunque no se nombre o se especifique que se están utilizando los datos no interpolados de la fase de preprocesado, se asume que se va a trabajar con los datos no interpolados.

4.3.2 Correlación

A continuación se pasa a estudiar cómo se relacionan las 3 variables principales del estudio: potencia, radiación y temperatura; utilizando los datos no interpolados de la fase de preprocesado.

La relación que tienen estas variables se puede estudiar midiendo la correlación que existe entre cada par de variables. Es decir, se va a estudiar la correlación de la potencia con la radiación ("power-poa"), la potencia con la temperatura("power-Tcell") y la radiación con la temperatura ("poa-Tcell").

En Python se puede utilizar la función "*corr()*" sobre el *dataframe* que se quiera analizar la correlación de sus columnas. Así se ha hecho, obteniendo los resultados que se pueden observar en la Tabla 15.

	power-poa	*power-Tcell*	*poa-Tcell*
Correlación df1_ni	*0.991012*	*0.797953*	*0.829644*
Correlación df2_ni	*0.989487*	*0.809163*	*0.834032*
Correlación df3_ni	*0.989372*	*0.804261*	*0.833197*
Correlación df4_ni	*0.988035*	*0.801035*	*0.832755*

Tabla 15. Correlación no interpolado

La función de correlación se ha ejecutado para los 4 *dataframes* para obtener las correlaciones en los distintos campos solares. Y se ha observado que entre campos solares no cambian las relaciones que existen entre las variables.

Si la correlación entre dos variables es cercana a cero indica que esas variables no presentan ninguna relación entre ellas. En cambio, cuando la correlación va aumentando y se aproxima a uno, las variables presentan mucha relación entre ellas. Cuando la correlación entre dos variables es igual a uno significa que las variables se

comportan exactamente igual y siguen la misma tendencia. Esto es prácticamente imposible con las magnitudes del estudio. La única correlación que va a ver igual a uno es entre una variable y ella misma (Ej.: "power-power").

En la Tabla 15, observando los resultados de correlación de cualquier campo solar, se puede concluir que la correlación más alta es entre la potencia y la radiación, con un valor aproximadamente de 0.99. La siguiente correlación más alta después de la ya mencionada es la de la radiación con la temperatura, que tiene un valor de 0.83, y por último la menor es la correlación entre la potencia y la temperatura con un valor de 0.80.

Aunque las dos últimas correlaciones: la de la temperatura con la radiación y la temperatura con la potencia, son muy similares.

Una forma de visualizar las correlaciones de manera gráfica es mediante los mapas de calor. Donde cada correlación corresponde con un color. En este caso, el color más suave corresponde con correlaciones altas y los colores más oscuros con correlaciones bajas. Se ha representado en la Figura 47 un mapa del calor del campo solar 1 en el que se puede visualizar las correlaciones.

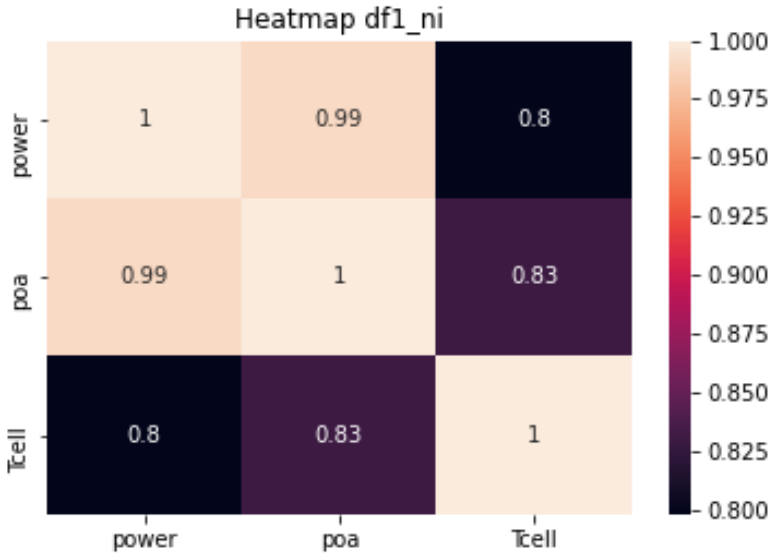

Figura 47. Mapa de calor df1

La correlación de la potencia con la temperatura es menor que la correlación de la potencia con la radiación porque la temperatura tarda más en ser detectada que la radiación, tras sufrir cambios. Esta diferencia de correlación se debe a que el sensor de radiación capta instantáneamente los cambios de energía solar, sin embargo el sensor de temperatura capta los cambios de forma más paulatina.

Aunque las correlaciones aporten una idea de cómo se relacionan las variables, es muy interesante visualizar esta relación o tendencia mediante gráficas. Por lo que para una mejor interpretación visual se va a implementar una función a la que se le indique: una variable, un día y un campo solar, para que esta función represente como se comporta la variable en ese día concreto.

En la Figura 48 se puede apreciar como evoluciona la potencia el 22 de Mayo de 2014. Aunque no se puedan apreciar muy bien los valores de los ejes, en el eje "x" se encuentran los valores de "Timestamp" desde las 6:15 horas de la mañana a las 20:15 horas de la noche. No se encuentran todos los datos del día ya que eliminamos valores con potencia nula; así que, es normal que no haya valores de potencia durante la noche. Y como se borraba todas las filas con potencia igual a cero, la radiación y la temperatura va a tener el mismo rango horario y por lo tanto el mismo eje "x" en la Figura 49 y en la Figura 50.

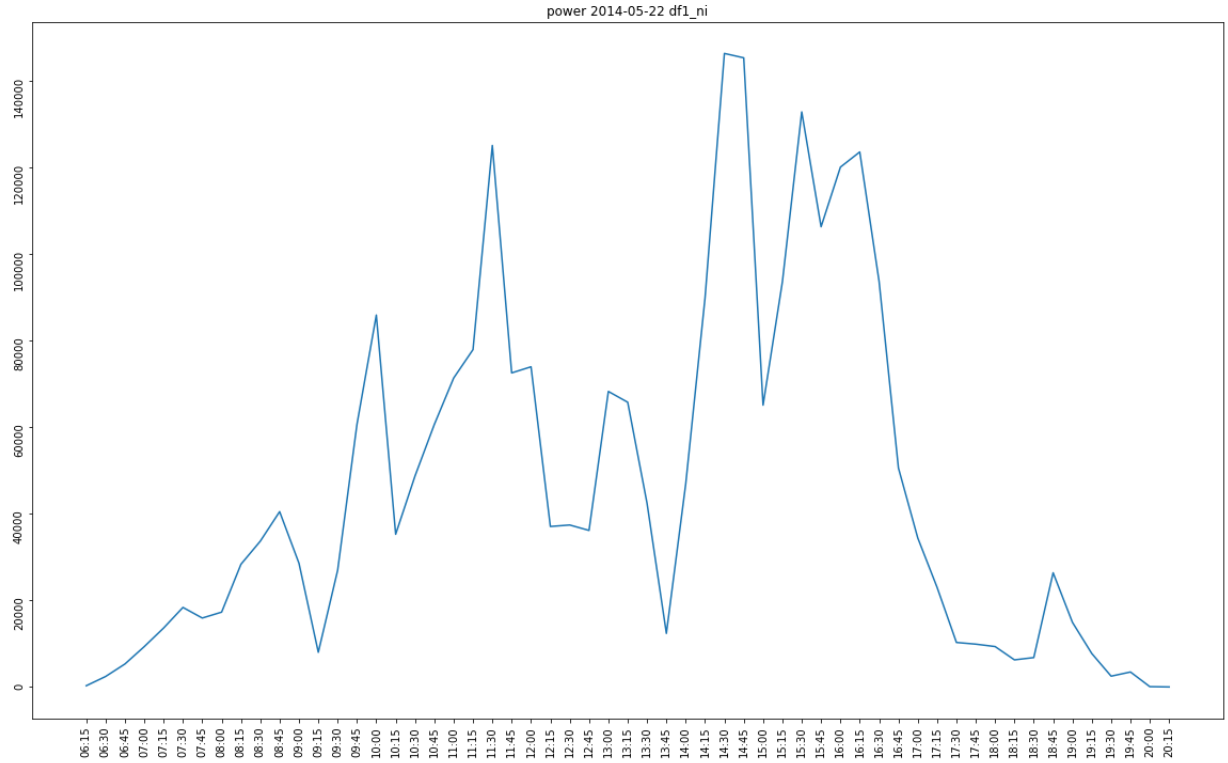

Figura 48. Potencia el 22 de Mayo de 2014

Aunque el eje x sea igual para las 3 variables, los valores del eje "y" no van a ser los mismos, ya que cada variable tiene un rango diferente. Pero lo importante es quedarse con la tendencia de las gráficas en las distintas variables para poder compararlas y comprobar si existen relación entre ellas.

Si se visualiza la Figura 48, se puede entender que existe menor potencia en las primeras horas de la mañana y a últimas horas de la tarde. Y durante el medio día, entre las 11:30 horas y las 16:30 horas es donde mayores picos de potencia se encuentran.

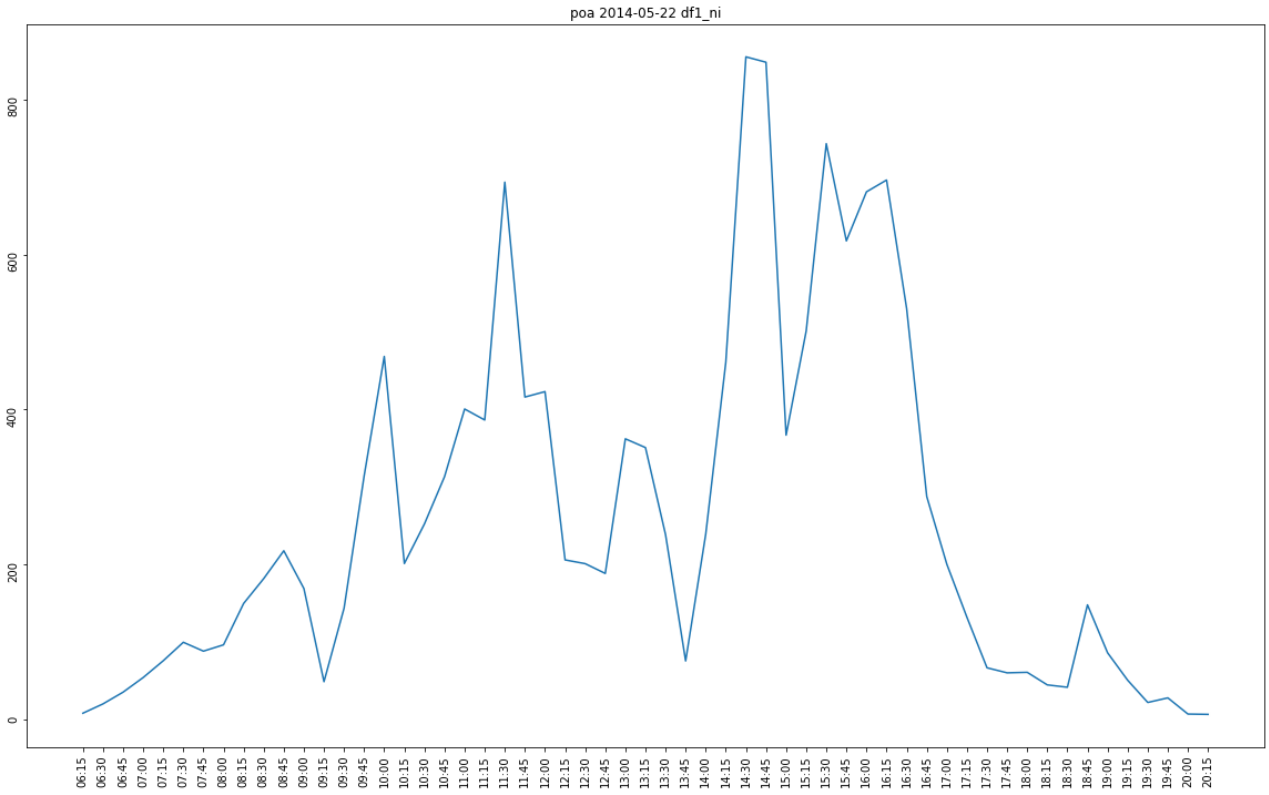

Figura 49. Radiación el 22 de Mayo de 2014

Esta misma tendencia se observa en la Figura 49 (que representa la radiación) y en la Figura 50 (que representa la temperatura), ya que las horas con mayor radiación solar y mayor temperatura suele ser a medio día. Y al amanecer y al atardecer se reduce la radiación y la temperatura, tal y como se ve reflejado en el comportamiento de la potencia. Este efecto demuestra que tanto como la radiación como la temperatura son factores claves para determinar la potencia.

Aunque de las tres representaciones, las gráficas que más se parecen son la de la potencia y la radiación como era de esperar. Pero igualmente la temperatura sigue la misma tendencia que la potencia.

En la representación de la temperatura se puede verificar lo afirmado anteriormente de que la temperatura no tiene cambios muy bruscos. Obviamente la temperatura tiene picos porque va cambiando a lo largo del día pero estos picos no son tan pronunciados como los de la radiación y la potencia.

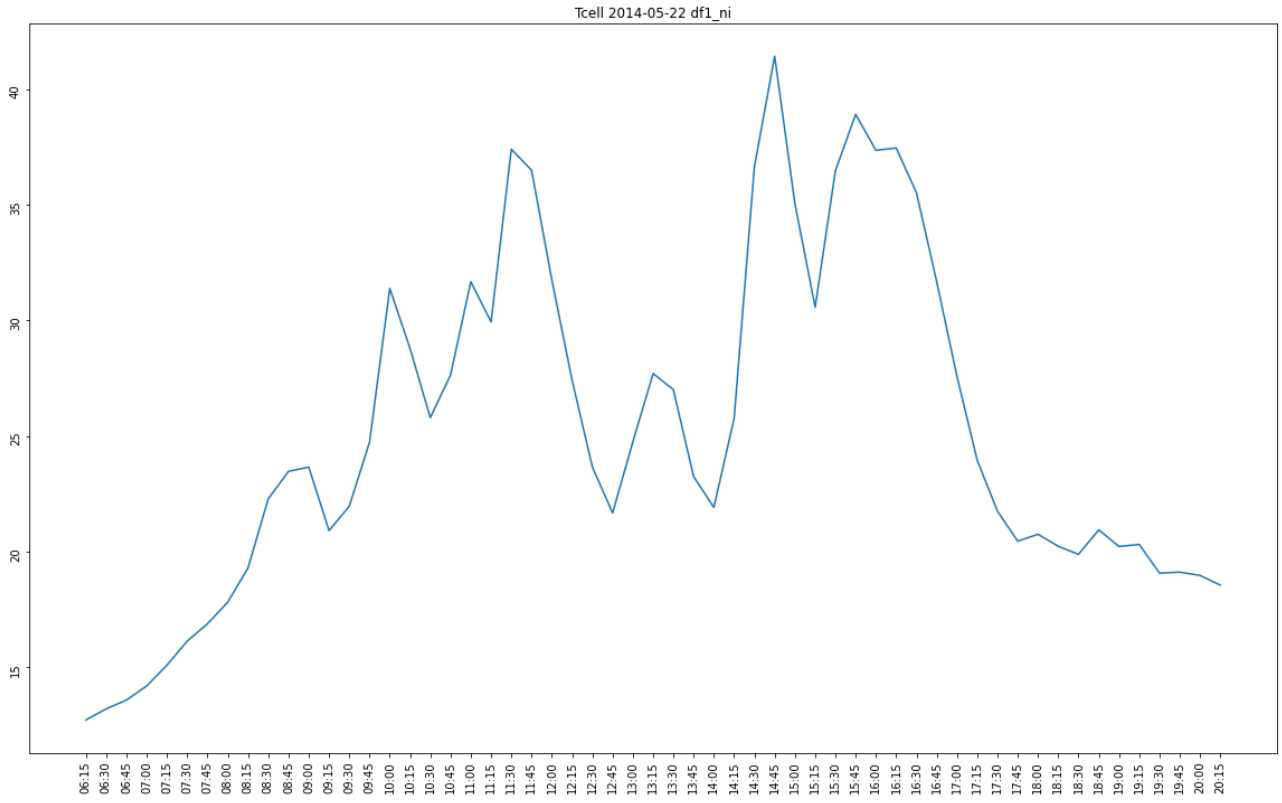

Figura 50. Temperatura el 22 de Mayo de 2014

Al no tener las 3 gráficas superpuestas, es más complicado comparar si la tendencia es exactamente la misma en ese instante de tiempo para las 3 variables. Pero no se pueden superponer estas gráficas porque el eje "y" es diferente para cada variable ya que la potencia tiene un rango, la radiación otro y la temperatura otro.

La solución ante este problema, es normalizar las variables para poder representar la potencia, la radiación y la temperatura en una misma gráfica y visualizar mejor como se relacionan entre ellas. Por lo que, las tres variables van a ser normalizas con la función "*MinMaxScaler*" de la librería "*sklearn*", pudiendo así transformar las variables. Estas variables son escaladas a un rango determinado, o el que viene por defecto que es entre 0 y 1. En este estudio, las 3 variables se han normalizado entre 0 y 1, y posteriormente se han representado juntas en la Figura 51.

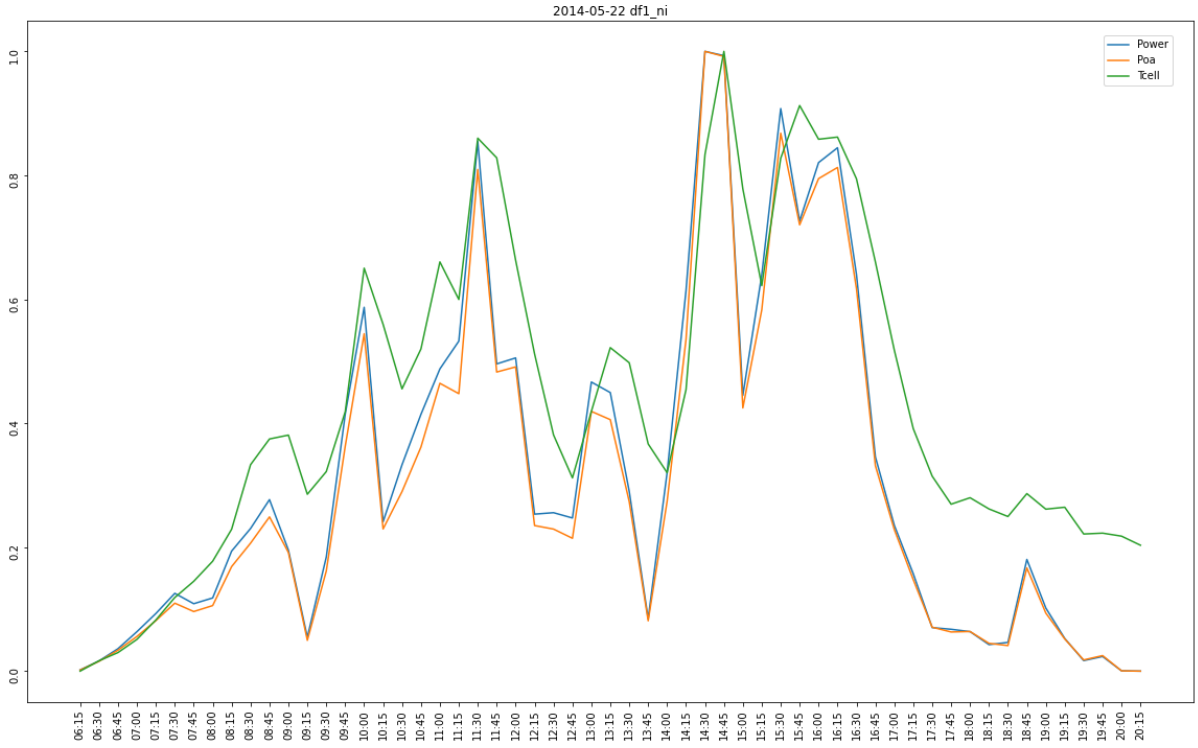

Figura 51. Variables normalizadas el 22 de Mayo de 2014

Ahora se pueden comparar mejor ya que la misma gráfica contiene las 3 tendencias. En la parte superior derecha de la Figura 51 se encuentra la leyenda, que indica que la línea azul representa la potencia, la naranja la radiación y la verde la temperatura.

La línea azul de la potencia y la línea naranja de la radiación son casi idénticas, sin embargo la línea verde de la temperatura se parece a estas dos pero no tanto. Entonces, lo que se había concluido anteriormente con las gráficas por separado, era cierto; existe una correspondencia entre las 3 variables, pero la de la potencia con la radiación es la más fuerte.

Uno de los objetivos de este estudio, aparte de estudiar la degradación del campo solar, es intentar realizar una predicción de la potencia solo con datos de radiación y temperatura. Es decir, se desea comprobar de manera teórica si una ubicación es la correcta para desplegar una planta solar. Ya que antes de desplegarla, con los costes que esto conlleva, se desea saber si se va a rentabilizar.

Este estudio podría determinar la potencia a largo plazo utilizando información de la radiación y de la temperatura. El experimento consistiría en desplegar sensores de radiación y temperatura en la ubicación que se quiere estudiar y tomar gran cantidad de medidas. Sabiendo la potencia predicha, se podría calcular los beneficios que se obtendrían si se desplegase esa planta.

Por eso es interesante saber cómo se relacionan las variables entre sí. ¿Qué variable es la que mejor predice la potencia? La correlación más alta es la de la potencia y la radiación.

¿Entonces, no se podría predecir la potencia utilizando la temperatura? Sí que se puede predecir la potencia utilizando la temperatura, pero esta correlación es menor que la que tiene con la radiación. Ya que con la radiación es de 0.99 y con la temperatura es de 0.8.

¿Y no sería mejor utilizar tanto radiación como temperatura para predecir la potencia? Pues eso se va a estudiar en la siguiente fase del estudio. Pero lo que si que se sabe cierto es que si solo se pudiese predecir la potencia mediante una variable, teniendo que elegir entre radiación y temperatura, se elegiría la radiación ya que esta es la que más relación tiene con la potencia.

Capítulo 5 - Aprendizaje automático

En el capítulo 4 se ha realizado un tratamiento y análisis de datos para conseguir unos datos limpios que puedan ser utilizados en el aprendizaje automático. En ese estudio práctico se ha concluido que los datos que se deben utilizar en este estudio son los datos no interpolados.

A continuación en este capítulo se va a continuar con la otra parte del estudio práctico que trata sobre el aprendizaje automático.

En el quinto capítulo se va a predecir el rendimiento de una planta solar utilizando modelos de regresión. Estos modelos van a ser: regresión lineal y bosque aleatorio. Cabe destacar que en ambos modelos se utilizan algoritmos de regresión y no de clasificación.

Y posteriormente, en este capítulo se va a estimar el porcentaje de degradación anual que han experimentado los paneles solares.

5.1 Regresión

Ahora que ya se ha realizado la parte de preprocesado de los datos y se ha hecho un análisis estadístico de estos datos; se sabe qué datos se deben utilizar, la cantidad de estos, cómo se distribuyen, cómo se relacionan entre ellos, etc.

Como se ha concluido en la fase anterior del estudio, los datos que se van a utilizar son los datos no interpolados. Cada campo solar va a seguir teniendo un *dataframe* independiente que está compuesto por 4 columnas ('Timestamp', 'power', 'poa' y 'Tcell) y 112179 filas. Estas columnas al igual que en la fase anterior, van a quedar así tras ser renombradas, ya que la radiación y la temperatura se guardaron como 'media_pa' y 'media_temp'.

Como bien se ha explicado al final de la fase anterior; uno de los objetivos que tiene este estudio, es el de poder predecir la potencia utilizando sensores de radiación y temperatura. Este experimento se hace para conocer teóricamente si en esas condiciones climatológicas la planta solar puede obtener suficiente potencia como para generar los beneficios deseados.

Por lo tanto, en esta fase de aprendizaje automático se van a realizar modelos de aprendizaje supervisado para deducir la potencia a partir del resto de las columnas con dos algoritmos:

- Regresión lineal ("*Linear Regression*" en inglés, que en el código implementado se abrevia a veces a "lr")
- Regresión de bosques aleatorios ("*Random Forest Regressor*" en inglés, que en el código implementado se nombra en ocasiones como "rf")

En primer lugar, se va a explicar la regresión lineal y como se ha implementado sobre los datos para predecir la potencia. Como para la regresión de bosques aleatorios es necesario entender previamente los árboles de decisión, se van a introducir los conceptos de forma teórica en el siguiente apartado, después de la regresión lineal.

Una vez entendido los arboles de decisión se van a explicar los modelos de bosques aleatorios y también como se ha diseñado este modelo predictivo sobre los datos de este estudio, y cómo se pueden optimizar los hiperparámetros para obtener mejores resultados.

Se va a intentar predecir la potencia: utilizando solo la radiación, utilizando solo la temperatura, y utilizando tanto radiación como temperatura. Son 3 casos distintos que van a ser probados tanto para la regresión lineal como para los bosques aleatorios. Esto es equivalente a decir que se van a implementar 6 modelos: 3 modelos de regresión lineal y 3 modelos de bosques aleatorios.

Se trata de generar un modelo por cada caso explicado: un modelo que predice la potencia utilizando solo una variable (la radiación), un modelo que predice la potencia utilizando solo una variable (la temperatura) y un modelo que predice la potencia utilizando dos variables (radiación y temperatura). Esto se va a implementar para la regresión lineal y para los bosques aleatorios.

Para los modelos que predicen con los datos de una sola variable no es necesario diseñar dos códigos distintos en Python, sino que se puede hacer mediante una función a la que se le pase la variable con la que se desea predecir. Así se disminuye la cantidad de código.

Cada modelo va a realizar unas predicciones, las cuales tendrán unos errores dependiendo de cómo de bueno sea este modelo. Por lo que se van a tener 6 resultados distintos de RMSE ("*Root Mean Squared Error*", que en castellano se traduce como "error cuadrático medio"). Es decir, se obtienen tres errores para cada uno de los modelos. Este error cuadrático medio es una medida del error que existe entre los valores predichos y los valores reales medidos. Entonces, cuanto menor sea este error, mejor van a ser las predicciones del modelo, que es lo que se quiere.

Y por último, se va a realizar una comparativa de los 6 modelos distintos, en base al RMSE y se va a decir cual es el que mejor predice.

5.1.1 Regresión Lineal

Los modelos de regresión lineal son algoritmos de aprendizaje supervisado. Los modelos de aprendizaje supervisado son aquellos que aprenden funciones (relaciones que asocian las entradas con las salidas). Por lo que estos modelos se ajustan ("*fit*") a un conjunto de ejemplos se le pasan. Estos ejemplos que se otorgan al modelo, son muestras de entrenamiento. Es decir, se conoce tanto la entrada como la salida y es para que el modelo entrene con esos datos.

Posteriormente los modelos de aprendizaje con lo que han aprendido de los ejemplos (entrenamiento) van a intentar estimar la salida cuando le demos una entrada que no está contemplada en los ejemplos. Pero esto no implica que esa estimación esté correcta. Se trata de una predicción en base a lo aprendido y por lo tanto se puede equivocar.

La regresión lineal consiste en encontrar la relación lineal entre varias variables. La idea principal de este algoritmo es conseguir una recta que se ajuste lo mejor posible a los datos. La recta que mejor se ajusta es aquella cuyo error de predicción total (todos los puntos de datos) es lo más pequeño posible. Este error es la distancia entre el punto y la línea de regresión. Se hace la media de las distancias de todos los puntos a la recta y ese es el error total del modelo.

El algoritmo de regresión lineal busca minimizar la función de coste y cuando esto se consigue y se traza la recta, esta es la recta óptima. Esta recta está compuesta por unos coeficientes óptimos que consiguen minimizar la función de coste.

La regresión lineal se utiliza para encontrar la relación lineal entre la variable objetivo y uno o más predictores. En concreto, busca la relación entre varias variables continuas.

En este caso, la variable objetivo es la potencia, que es la variable que se quiere predecir (variable independiente). Y el predictor es la variable que se va a utilizar para predecir la variable objetivo. El predictor es la variable dependiente.

Como se ha mencionado puede haber una o más variables que predigan la variable objetivo. Por eso se divide la regresión lineal en dos tipos según si la predicción es con una única variable o con varias. Si se predice con una variable se trata de regresión lineal simple, y si se predice con varias variables se trata de regresión lineal múltiple.

5.1.1.1 Regresión lineal simple

En este estudio se va a implementar una función que permita predecir la potencia usando un modelo de regresión lineal simple.

Antes de nada, es interesante saber si las variables continuas de radiación y temperatura tienen una relación lineal con la potencia, para saber si los datos de estas variables sirven para crear un modelo de regresión lineal. Así que se va a representar la relación entre potencia y radiación y la relación entre potencia y temperatura para un número determinado de puntos.

Esta función se ha implementado con el objetivo de realizar una representación de la relación entre dos variables, para ver si los puntos representados tienen forma lineal. Se va a utilizar la función de *Pandas "sample()"* para generar una muestra aleatoria de los datos y así no tener que representar todos los datos.

En la Figura 52 se muestra una representación de la relación entre la potencia y la radiación utilizando 500 puntos aleatorios. Se puede contemplar perfectamente que

estos puntos están distribuidos de forma lineal, por lo que podría pasar una recta entre ellos.

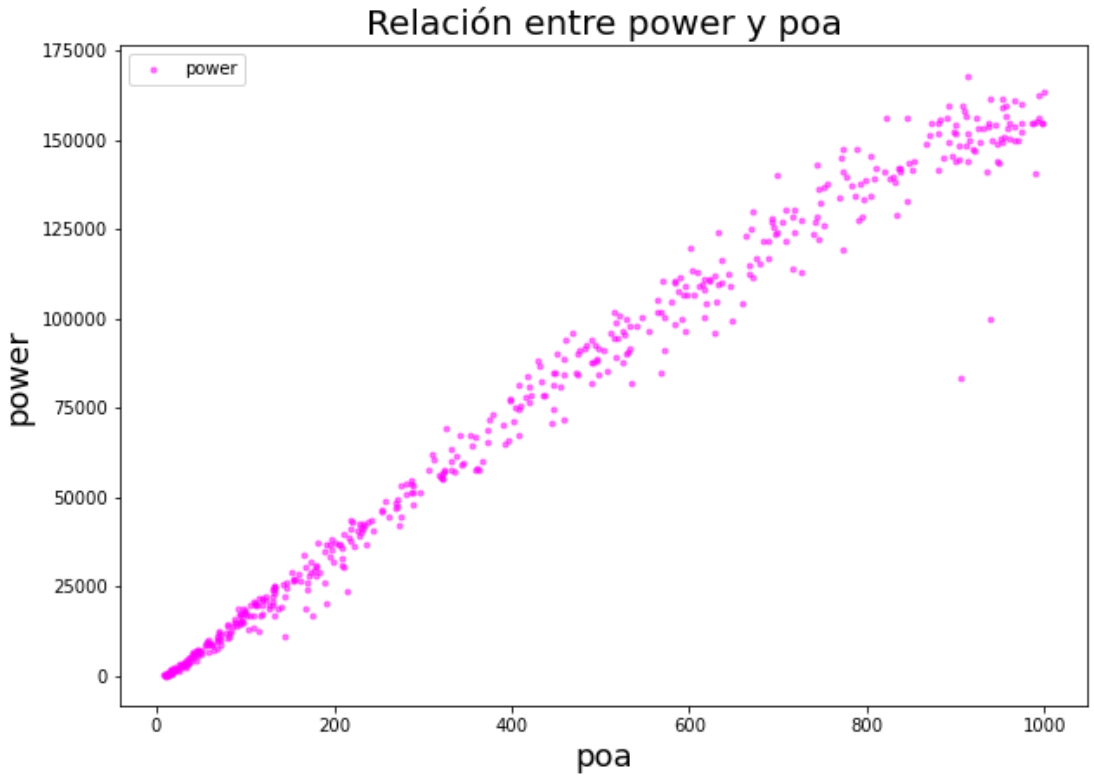

Figura 52. Relación lineal entre potencia y radiación

Y ahora en la Figura 53 se realiza la representación de la relación entre la potencia y la temperatura utilizando también 500 puntos aleatorios. Se puede apreciar que estos puntos están distribuidos de forma más dispersa que los puntos de la relación entre potencia y radiación, pero igualmente podría pasar una recta entre estos puntos. Algunos puntos quedarán más alejados de la recta de predicción y esos puntos serán los que más error aporten. Cabe esperar, que el error de la predicción de la potencia a través de la temperatura sea mayor que el error al predecir la potencia con la radiación.

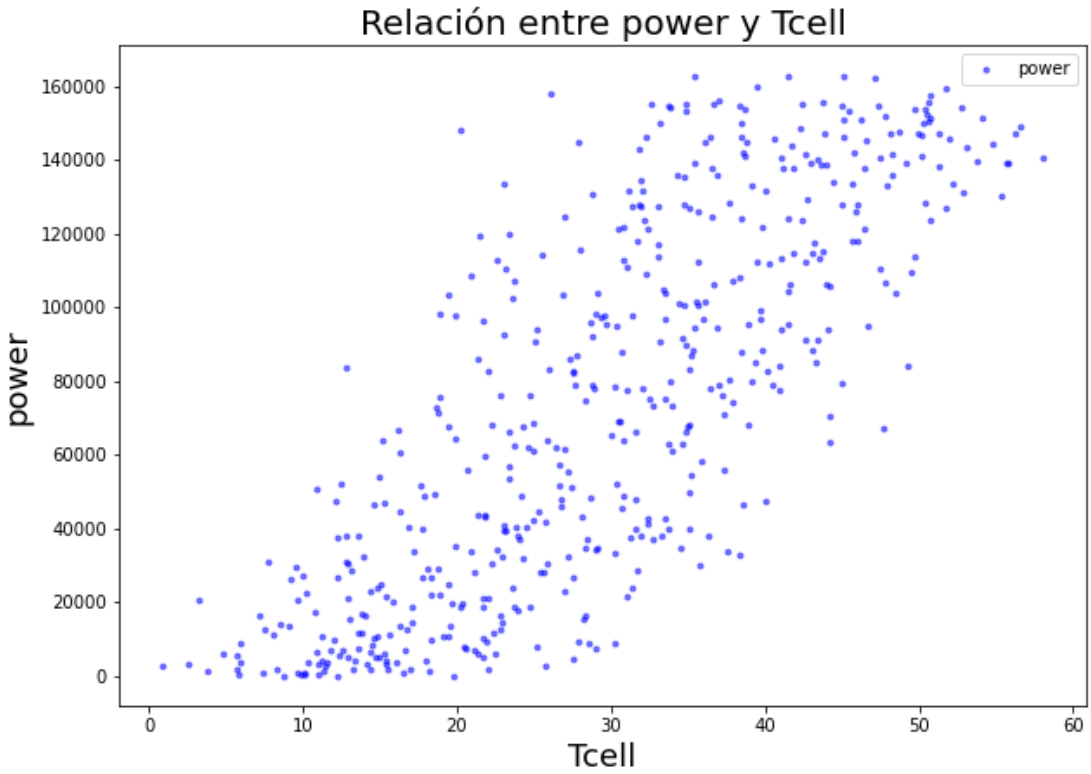

Figura 53. Relación lineal entre potencia y temperatura

A la función que se ha implementado para realizar los dos modelos de regresión lineal simple, se le pasa como parámetro la variable (radiación o temperatura) que va a ser la que va a predecir la potencia.

En esta función se selecciona la columna de datos de la variable que se ha especificado y se la nombra como "x", y por otra parte se selecciona la columna de la potencia como "y". A continuación se van a utilizar los datos "x" y "y" para dividir los datos en 4 grupos: 'x_train', 'y_train', 'x_test' y 'y_test', utilizando la función "*train_test_split*" la cual permite dividir los datos aleatoriamente en entrenamiento("train") y prueba ("test"). En este estudio se ha elegido que los datos dedicados para entrenamiento debían ser un 70% de los datos totales y un 30% de todos los datos iban a ser dedicados para pruebas.

Una vez se tienen los datos divididos para trabajar con ellos, se crea el objeto de regresión lineal. Seguidamente se entrena el modelo, pasándole al objeto que hemos creado los datos de entrenamiento, es decir: 'x_train' y 'y_train'.

Posteriormente se realiza la predicción de los datos. Esta va a ser la recta predicha, la cual se conoce como "y_pred". Esta recta es generada utilizando la función "*predict*" sobre el modelo, a la cual se le pasan los datos de entrenamiento: 'x_train'. Esta recta predicha es la que ha generado el modelo, y es la que intenta pasar por todos los puntos de entrenamiento que se le ha dado al modelo.

El algoritmo de regresión lineal busca minimizar la función de coste y cuando esto se consigue y se traza la recta, esta es la recta óptima. Esta recta está compuesta por un coeficiente óptimo que consigue minimizar la función de coste:

La recta predicha de la regresión lineal simple tiene la siguiente forma:

$$Y = m X + b$$

Esta recta está compuesta por solo 1 coeficiente: m y el termino independiente b al tratarse de regresión lineal simple. El coeficiente óptimo y el termino independiente pueden ser obtenidos del modelo para saber cual es la pendiente (m) y el origen (b) de la recta que ha predicho el modelo del estudio.

Mediante la radiación se ha predicho la siguiente recta de potencia que se observa en la Figura 54, que tiene los parámetros óptimos: m = 168.577 y b = 929.12.

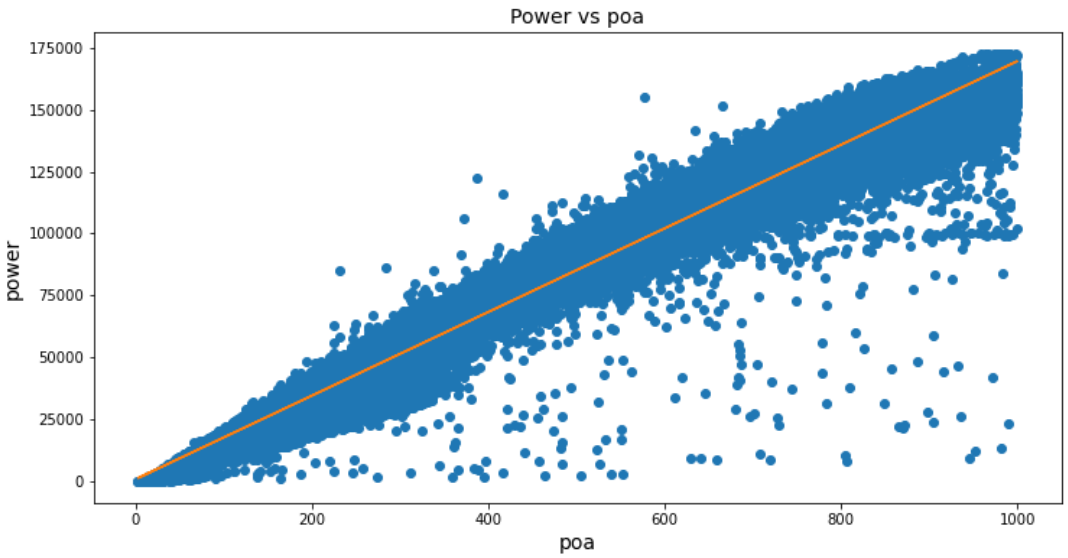

Figura 54. Predicción de potencia mediante radiación

Y mediante la temperatura se ha predicho la siguiente recta de potencia que se observa en la Figura 55, que tiene los parámetros óptimos: m = 3355.21 y b = -26551.23.

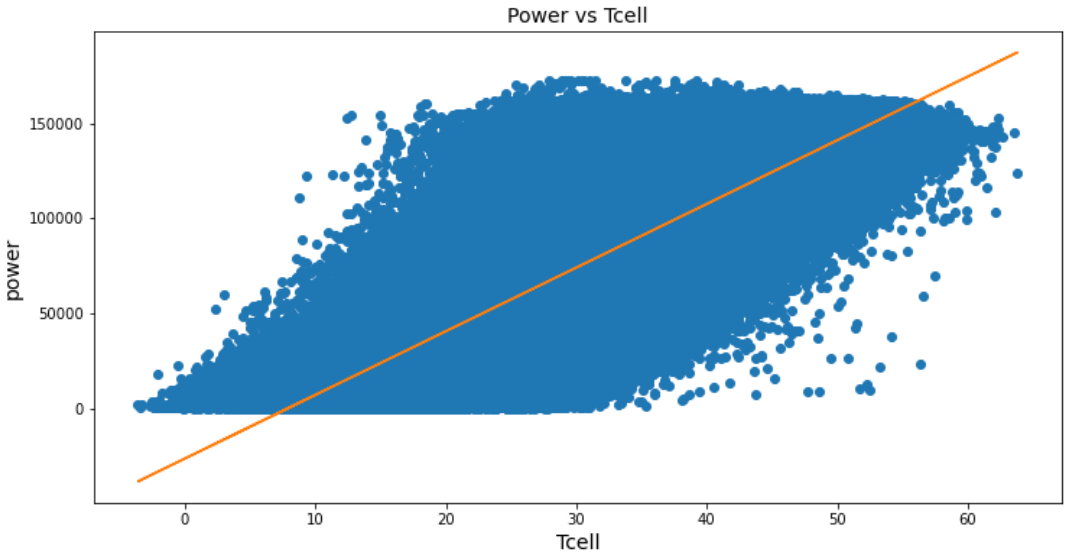

Figura 55. Predicción de potencia mediante temperatura

A continuación se va a evaluar el modelo. Ahora que se tiene la recta predicha se puede calcular la distancia que hay de los valores reales de potencia a esa recta (potencia predicha). Esta distancia se trata del error cuadrático medio o MSE, el cual se puede calcular utilizando la función "*mean_squared_error*". Esta función utiliza 'y_train' y 'y_pred' para calcular el error.

Solo hace falta aplicarle la raíz cuadrada al MSE para obtener el RMSE; y para esto se va a utilizar la función de *Numpy* "*sqrt*".

El RMSE obtenido para la radiación es de: 7035.6267 W y para la temperatura es de: 32075.0453 W Como se había anticipado, el error de la predicción de la temperatura iba a ser mayor al de la radiación.

Como el objetivo es intentar implementar un modelo que prediga la potencia lo mejor posible; si hubiese que elegir entre predecir con radiación o temperatura, se elegiría el modelo de regresión lineal simple de radiación.

¿Qué quieren decir estos valores de RMSE? Que el valor real de la potencia puede estar un rango determinado desviado de la potencia predicha. Es decir, el valor

real de la potencia es igual al valor predicho ± rango de error en el 95% de las veces. Ese rango de error es dos veces el RMSE.

Por lo tanto:

valor real de la potencia = valor predicho de la potencia ± 2 x RMSE,

en el 95% de los casos.

Por lo que, si se hiciese una predicción de la potencia con la radiación se obtendría que:

Valor real de la potencia = valor predicho de la potencia ± 2 x 7035.6267 W,

Valor real de la potencia = valor predicho de la potencia ± 14071.2534 W

Este error puede estar por encima o por debajo de la recta, por lo que se utiliza el ± . Así que el valor real se va a encontrar en el valor predicho con un rango de error de 14071.2534 W por encima y por debajo.

Otra de las medidas que se utiliza para saber lo bueno que es un modelo es la correlación R^2, que indica la relación entre dos variables. Se suele utilizar R^2 porque es más fácil de interpretar. R^2 es el porcentaje de variación (varía de 0 a 1) explicado por la relación entre dos variables.

La varianza de todo el conjunto de datos es igual a la suma de la distancia entre cada punto de datos y la media, todo al cuadrado. La diferencia se eleva al cuadrado de forma que los puntos por debajo de la media no se anulen con los puntos por encima de la media.

Varianza(media) = suma(pi - media)²

Suponiendo que se utiliza la regresión lineal para encontrar la recta que mejor se ajusta. El valor de R^2 se puede expresar entonces como:

R^2 = (var(media) - var(recta)) / var(media)

Donde var(media) es la varianza con respecto a la media y var(recta) es la varianza con respecto a la recta.

Este valor de R^2 ha sido de un 0.98256 para la radiación y de un 0.63769 para la temperatura. Como se desea que R^2 sea lo más cercano a uno para un buen modelo,

se puede comprobar también con esta otra medida que la radiación es mejor predictor que la temperatura.

El valor de R^2 implica que hay un 0.98256% menos de variación alrededor de la recta que de la media. En otras palabras, la relación entre la potencia y la radiación explica el 0.98256% de la variación. Dicho de otra manera, la radiación es un buen predictor de la potencia porque cuando la radiación aumenta también lo hace la potencia y viceversa.

Tanto el RMSE, como el R^2 son las dos medidas que se realizan para evaluar el modelo y poder concluir si es un buen modelo. Y con estas medidas finaliza el proceso de predecir la potencia utilizando la regresión simple (mediante la radiación o la temperatura). Así que se podría concluir, indicando de forma más resumida y esquemática los pasos que se han seguido en esta explicación:

1. Dividir los datos en datos de entrenamiento y datos de prueba.
2. Generar el modelo y entrenarlo con los datos de entrenamiento
3. Evaluar el modelo con los datos de test
4. Calcular RMSE (Evaluar el modelo)
5. Calcular R^2 (Evaluar el modelo)

Cabe destacar que este proceso de dividir datos, entrenar el modelo y evaluarlo no se ha repetido una sola vez. Este proceso que se ha resumido en 5 fases se ha repetido 30 veces y se ha hecho la media del RMSE y el R^2 para asegurarse de que se estaba evaluando el modelo de una manera fiable. Ya que cuando se trabaja con datos que son seleccionados aleatoriamente, el resultado de las medidas RMSE y R^2 nunca van a ser iguales. Este mecanismo para asegurarse de que el RMSE y R^2 que se obtiene es fiable se implementa para todos los modelos; tanto para regresión lineal como para bosques aleatorios.

5.1.1.2 Regresión lineal múltiple

Y por otro lado, se va a diseñar un modelo de regresión lineal múltiple que permita predecir la potencia a través de la radiación y de la temperatura. En este caso se utilizan dos predictores en vez de uno, como ocurría antes en la regresión lineal simple.

Ahora en la regresión múltiple, se va a tener el mismo número de coeficientes que predictores. Por lo tanto, la recta va a estar compuesta de dos coeficientes más el termino independiente, ya que existen dos predictores (radiación y temperatura) para predecir la potencia.

Entonces la recta va a tener la siguiente forma:

$Y = b + m1\ X1 + m2\ X2$

Donde Y va a ser la potencia predicha, b el termino independiente, X1 es la radiación y X2 es la temperatura, y m1 y m2 son los dos coeficientes óptimos.

Como se puede apreciar, ahora se obtienen 2 coeficientes (cada uno correspondiente a una de las 2 variables predictivas). Así que, ahora ya no se puede graficar una recta si no que tiene que ser en un plano de 3 dimensiones.

Básicamente el procedimiento a seguir para crear un modelo de regresión lineal múltiple es el mismo que para el de regresión lineal simple. Lo único que ahora a la función que crea el modelo no se le pasa el nombre de la variable como parámetro para que seleccione esos datos. Por lo tanto dentro de la función implementada el valor de la variable "x" es una combinación de las dos columnas: radiación y temperatura. Por lo que "x" ahora tiene un tamaño de: (112179, 2) y antes era de (112179, 1). Y la "y" sigue siendo igual que antes, es la columna de la potencia.

Por lo demás, se sigue el mismo procedimiento: se dividen los datos utilizando la "x" y la "y" definidas, se crea el modelo, se entrena y se evalúa utilizando las medidas de RMSE y R^2.

Y tras ejecutar el bucle que itera 30 veces y hace la media de RMSE y R^2, los resultados obtenidos son los siguientes:

RMSE= 6634.0261W

$R^2 = 0.98450$

Y los coeficientes obtenidos que minimizan la función de coste son: m1=179.47678708 y m2=-324.93271842; y el termino independiente es b= 5828.86.

Como se ha podido comprobar, el RMSE utilizando la radiación y la temperatura a la vez para predecir la potencia es el menor de todos y el R^2 es el mayor de todos; que es exactamente lo que nos índica que se trata de un buen modelo. Por lo que si se tuviese que elegir si utilizar regresión lineal simple o múltiple para predecir la potencia, se elegiría la múltiple ya que con ambas variables se logra predecir mejor la potencia.

5.1.2 Árbol de decisión

El algoritmo de árbol de decisión es un método de aprendizaje supervisado capaz de encontrar relaciones no lineales y más complejas en los datos. Este algoritmo puede realizar tanto regresión como clasificación. Pero para el estudio el algoritmo que resulta interesante es el de regresión. Se utiliza entonces la clase "*DecisionTreeRegressor*" de "*sklearn*".

Los árboles de decisión son modelos que utilizan varias reglas binarias del tipo "si" o "no", para clasificar los resultados. Cada árbol individual es un modelo simple que está compuesto por ramas, nodos y hojas [26]. Este algoritmo se denomina árbol de decisión porque tiene una estructura en forma de árbol, como se puede observar en la Figura 56. Este árbol dispone de tres tipos de nodos: el nodo raíz conocido como "*root*", el nodo interior y el nodo hoja conocido como "*leaf*". El algoritmo tiene esta estructura de árbol pero del revés; es decir, el nodo raíz está arriba y los nodos hoja están abajo.

Como se puede apreciar en la Figura 56, nodo raíz solo pude haber uno; se trata del nodo inicial y puede dividirse en otros nodos. Los nodos interiores representan las características de un conjunto de datos y las ramas representan las reglas de decisión, también conocidas comúnmente como condiciones. Y los nodos hoja son los resultados obtenidos tras clasificar los datos siguiendo las condiciones.

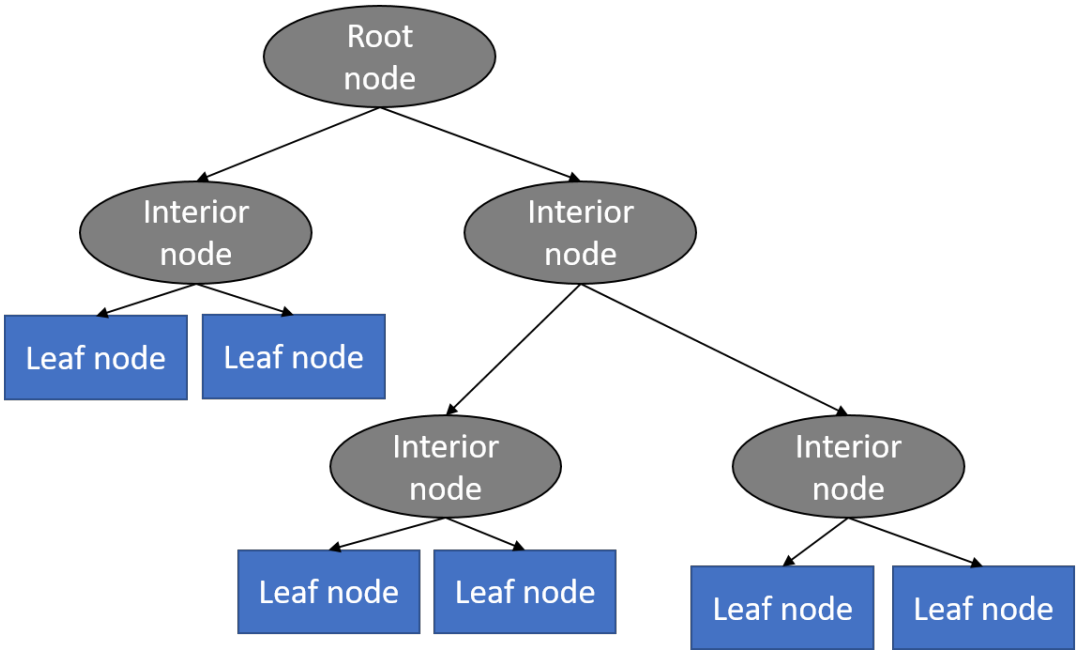

Figura 56. Estructura árbol de decisión [27]

Se pregunta a los datos si estos cumplen las condiciones establecidas y se van clasificando los datos dependiendo de las condiciones que satisfacen. Gracias a estas condiciones se llega a una estimación. Cada regla de decisión reduce los posibles valores. Entonces, cuanto más nodos existan, mejor clasificados van a estar los datos y mayor precisión va a tener modelo y va a realizar mejores predicciones. El orden de las preguntas y el umbral de estas condiciones, lo determina el modelo.

En todos los nodos, menos en los nodos hojas, existen estas preguntas o condiciones. Y de estas reglas de decisión salen dos ramas, una que indica que el dato satisface la condición, y otra que indica que el dato no satisface la condición.

Esto se puede visualizar en la Figura 57, donde "SI" significa que sí que satisface la condición y "NO" que no la satisface. Independientemente de las respuestas a las preguntas, al final se llega a una predicción que es el nodo hoja.

La clasificación de los datos se realiza de arriba hacía abajo (desde el nodo raíza hasta los nodos hoja). Es decir, se selecciona un valor del conjunto de entrenamiento y

se empieza por el nodo raíz en la parte superior y se avanza por el árbol respondiendo a las preguntas por el camino.

Esta clasificación de los datos es el entrenamiento, con el que se consigue que los datos estén distribuidos en zonas que están delimitadas por las condiciones. Entonces en cada zona habrá valores que cumplan las mismas condiciones, y por lo tanto los valores de una misma zona van a ser de similares características (no iguales porque se utilizan intervalos). Una vez se han clasificado todos los datos se hace la media de los valores de cada zona.

El algoritmo explora las características en un orden aleatorio. Si no se define el parámetro: "*max_features*", debería explorar todas las características y encontrar la mejor división.

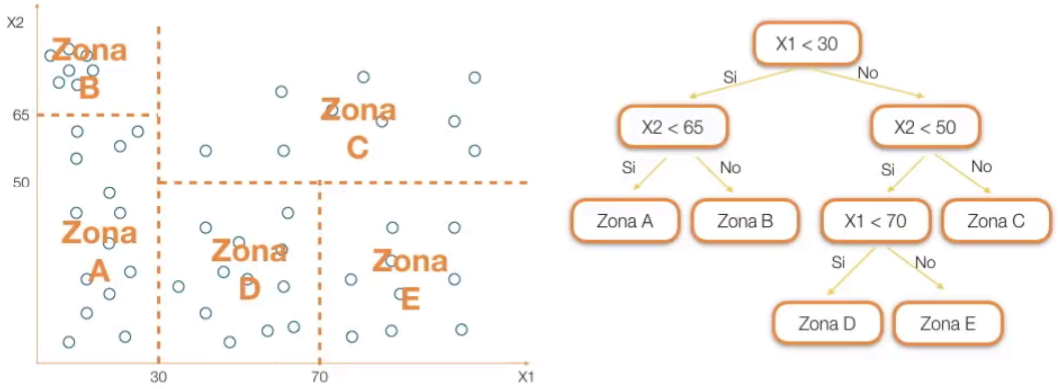

Figura 57. Árbol de decisión condición y zonas [28]

Los valores de las hojas terminales, que son las medias de cada zona, se utilizan para predecir el valor de cualquier nuevo valor. Se puede ver un ejemplo de un árbol de decisión muy simple en la Figura 58, donde se puede apreciar a la izquierda cómo se clasifican los datos según las condiciones y a la derecha, cómo se distribuyen los datos en zonas según las condiciones. Y en cada zona se ve un valor que es la media de todos los valores de la zona.

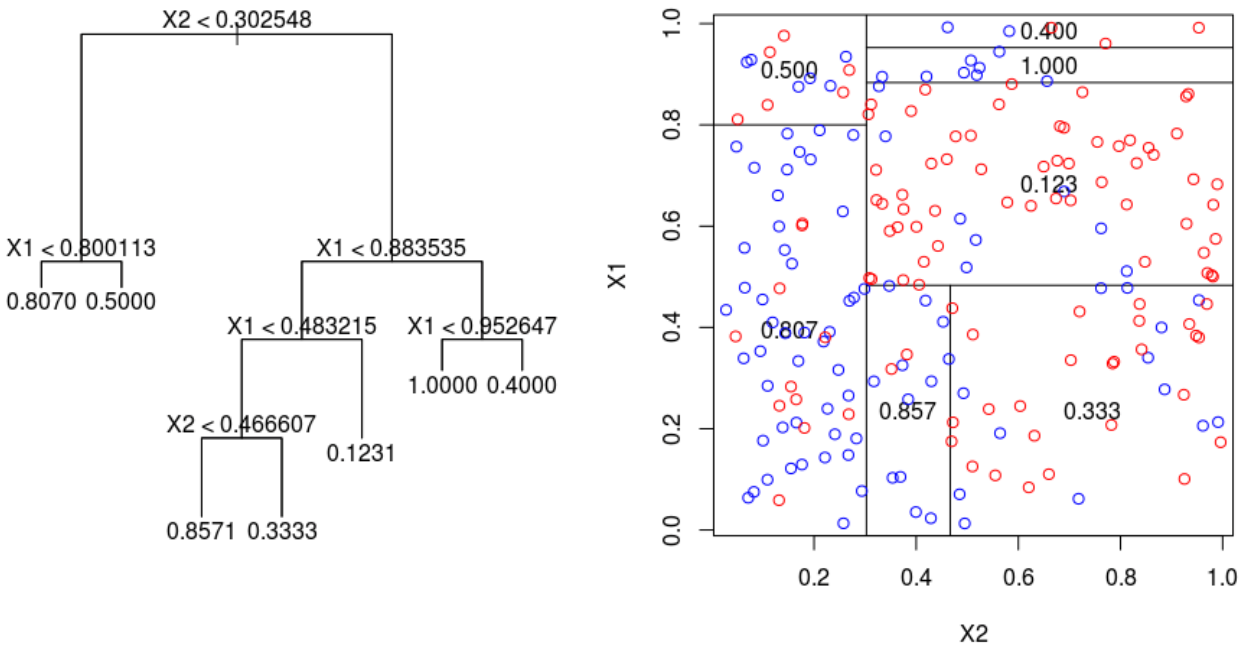

Figura 58. Medias de cada zona en un árbol de decisión [29]

El problema de los arboles de decisión es que si son muy simples (menor profundidad, es decir pocos nodos), crean modelos con alto sesgo, y sin embargo si son más complejos (mayor profundidad, es decir muchos nodos) son más propensos al sobreajuste y, por tanto una mayor varianza en el modelo. Este problema de los arboles de decisión es resuelto con el modelo de bosque aleatorio o "*Random Forest*" [30]. Un bosque aleatorio es un modelo formado por muchos árboles de decisión. Y utilizando varios arboles de decisión se puede mejorar la varianza global del modelo

5.1.3 Bosque aleatorio

Ahora que ya se entiende cómo son los árboles de decisión o "*Decision Tree*", tras haber visto un resumen teórico en el apartado anterior; se pasa a estudiar los bosques aleatorios, también conocidos como "*Random Forest*"; que estos si que van a ser usados en este estudio.

Un bosque aleatorio es un modelo construido mediante varios árboles de decisión. Como el modelo anterior se llamaba árbol de decisión por su estructura, este

modelo se llama bosque porque utiliza un conjunto de árboles. Se trata también de un algoritmo de aprendizaje supervisado, al igual que la regresión lineal. Los modelos de bosque aleatorio son muy útiles y son fáciles de usar ya que los parámetros predeterminados suelen producir buenos resultados. Como desventaja tiene que si se usa una gran cantidad de árboles puede hacer que el algoritmo sea lento.

Como el modelo consta de un conjunto de árboles de decisión, las predicciones de este modelo se realizan promediando las predicciones de cada árbol de decisión. Los resultados de las predicciones suelen ser mejor en estos modelos, ya que son más potentes que un único árbol de decisión.

El modelo de bosque aleatorio se suele usar para cuando existen relaciones no lineales o más complejas entre las características. Este es más robusto que un solo árbol de decisión, ya que utiliza un conjunto de árboles de decisión no correlacionados.

Cada árbol de bosque aleatorio se entrena con un conjunto de datos aleatorios, mientras que en cada nodo de decisión se considera un conjunto aleatorio de características.

Si se limita la profundidad máxima de un árbol de decisión, la varianza se reduce, pero el sesgo aumenta. Para conseguir que varianza y sesgo sean bajos, se necesita combinar muchos árboles de decisión con la aleatoriedad para reducir el sobreajuste. Y esto es lo que hace el bosque aleatorio.

Los datos de entrenamiento se dividen aleatoriamente generando pequeños subconjuntos de datos. Estos subconjuntos también se conocen como muestras *bootstrap* [31]. A continuación, estas muestras *bootstrap* se introducen como datos de entrenamiento en muchos arboles de decisión. Cada uno de estos arboles de decisión se entrena por separado con estas muestras *bootstrap*. El resultado final del modelo se obtiene realizando la media del resultado de cada árbol de decisión.

El "*Bagging*" es una abreviatura de "*Bootstrap Aggregation*", que es una técnica para reducir la varianza de un modelo de aprendizaje.

Bagging se consiste en extraen varias muestras aleatorias con reemplazo del conjunto de datos original para crear varios conjuntos de datos aleatorios más pequeños. En este caso, se realiza el *bagging* tras separar los datos de entrenamiento y

prueba. De los datos de entrenamiento se obtienen varios subconjuntos de datos aplicando *bootstrapping*, y cada subconjunto se va a utilizar para para entrenar un modelo, como se puede observar en la Figura 59.

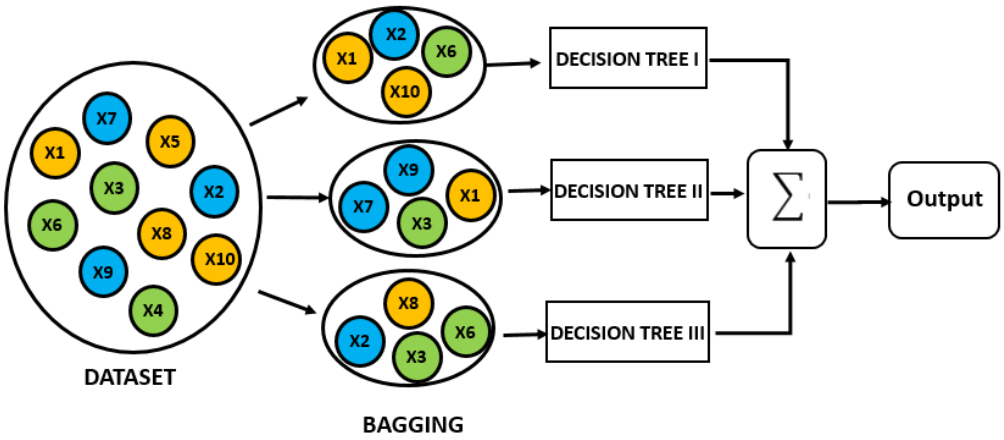

Figura 59. Bagging [32]

La agregación consiste en promediar los resultados de cada observación, a través de todos los modelos. Cada modelo se entrena en un conjunto de datos diferente. Así que, cada modelo cometerá diferentes errores y tendrá una varianza distinta. Tanto el error como la varianza van a ser promediados debido a la agregación, por lo que estos valores van a ser más pequeños.

A continuación se va a exponer en detalle como se crea el modelo de bosque aleatorio para este estudio:

1. Primero se separa el conjunto de datos en entrenamiento y prueba con los mismos porcentajes que en la regresión lineal. Y se tienen "M" características en ese conjunto de datos. Si "M" es uno, se trata de un modelo de bosque aleatorio simple, ya que solo se utiliza una variable para predecir la potencia y si "M" es dos, se trata de un modelo de bosque aleatorio múltiple, en el que se predice la potencia utilizando la radiación y la temperatura.
2. Se selecciona varios subconjuntos de datos de forma aleatoria del conjunto de datos de entrenamiento. Esta selección se realiza con

reemplazo; es decir, para la siguiente vez que se repita este primer paso de seleccionar datos, se van a seleccionar otros datos diferentes.

3. Se selecciona aleatoriamente un subconjunto de M características y se utiliza la característica que da la mejor división. En lugar de buscar la característica más importante al dividir un nodo, busca la mejor característica entre un subconjunto aleatorio de características.
4. Se repiten los pasos anteriores 2 y 3 de forma iterativa. La predicción del modelo se basa en la agregación de las predicciones de varios árboles.

Por lo general, se siguen unos pasos muy parecidos a los que se realizaron para los modelos de regresión lineal, se dividen los datos en entrenamiento y prueba, se entrena el modelo y se evalúa. En este caso solo se toma medida del RMSE, ya que no existe la medida de R^2 para los bosques aleatorios. R^2 describe de forma estadística cómo se ajustan las muestras a un modelo lineal. Y el modelo de bosque aleatorio no es un modelo lineal. Por lo que se va a utilizar el error cuadrático medio. Y además, se utilizará el error de "*Out of Bag*", que permite obtener una estimación del error de test. En los modelos de regresión de bosque aleatorio la métrica "*oob_score*" sería como el R^2 de un modelo lineal [33].

Cuando se produce *bootstrap*, es decir se crean subconjuntos del *dataset* original para dedicarlos a distintos arboles de decisión; no se cogen todos las muestras (ya que se trata de una selección aleatoria). Y por lo tanto hay muestras que no se quedan seleccionadas. Estas muestras que se quedan "fuera", se conocen como muestras "*Out of Bag*" y son diferentes para cada árbol de decisión. Como en cada árbol se seleccionan unas muestras distintas, cada árbol va a tener unas muestras que se queden sin seleccionar distintas.

Estas muestras que no han sido incluidas (muestras *OOB*) se entregan a los árboles de decisión como datos que no han contemplado para que predigan el resultado de estas muestras *OOB*. Y de esta predicción se obtiene el error de esas muestras, que se conoce como "*Out of Bag error*".

5.1.3.1 Bosque aleatorio simple

Al igual que con la regresión lineal, se ha querido predecir la potencia utilizando solo una variable (radiación o temperatura). Y como ocurría con la regresión lineal, la diferencia entre diseñar un modelo simple y uno múltiple solo radica en las columnas de datos que se seleccionan al principio. Ya que para bosques aleatorios se ha seleccionado como "x" la columna de la variable predictor (radiación o temperatura) y como "y" la columna de la variable predicha (potencia); siguiendo el mismo patrón que en la regresión lineal. Esta "x" y esta "y" son las dos columnas de datos que se va a pasar a la función "*train_test_split*" para que divida los datos en datos de entrenamiento y datos de prueba.

Se sigue la misma idea que para la regresión lineal. A la función implementada que crea el modelo, lo entrena y lo evalúa, se le pasa el nombre de la variable que va predecir la potencia.

Sin cambiar los hiperparámetros que se le pueden pasar al modelo, dejando los valores que ya vienen por defecto, se obtiene un RMSE al predecir con la radiación igual a 7315.1285 W, que es mayor que el RMSE en la regresión lineal con esta variable. Por lo que *random forest* simple crea un peor modelo con los parámetros por defecto.

Lo mismo ocurre con la temperatura, se obtiene un RMSE mayor en bosques aleatorios que en regresión lineal, ya que se obtiene un resultado de RMSE en bosques aleatorios igual a 32367.4289 W.

Por lo que se puede concluir, que si se utiliza solo una variable para predecir, y se utilizan los modelos de regresión lineal y bosques aleatorios con sus parámetros por defecto; es preferible utilizar la regresión lineal en ambos casos (tanto con temperatura como radiación)

5.1.3.2 Bosque aleatorio múltiple

Al igual que en los dos modelos de bosque aleatorio simple creados (uno para radiación y otro para temperatura) se han utilizado los valores por defecto que trae el modelo, para el bosque aleatorio múltiple también. Es decir, los resultados que se van a

mostrar en este apartado son los obtenidos para los modelos sin pasarles ningún parámetro. Y en el siguiente apartado, en la optimización de hiperparámetros se intentará mejorar el modelo.

Si se quiere predecir la potencia solo con una variable, se había comentado que era mejor utilizar la regresión lineal simple en vez de el bosque aleatorio simple.

En cambio para el bosque aleatorio múltiple en el que se utilizan ambas columnas: radiación y temperatura se obtiene un RMSE igual a 5632.1091 W, que es menor que el que se obtiene en regresión lineal múltiple: 6634.0261 W. Estos resultados se pueden comparar en la Tabla 1. Por lo tanto el error cuadrático medio, que nos indica lo bueno que es el modelo diseñado, dice que el error menor se obtiene cuando se utiliza la radiación (poa) y la temperatura (Tcell) a la vez para predecir la potencia. Y además este error es el más bajo en el bosque aleatorio.

	LR (poa)	*LR (Tcell)*	*LR (poa y Tcell)*	*RF (poa)*	*RF (Tcell)*	*RF (poa y Tcell)*
RMSE	*7035.6267*	*32075.0453*	*6634.0261*	*7315.1285*	*32367.4289*	*5632.1091*
R^2	*0.98256*	*0.63769*	*0.98450*	-	-	-
Error=RMSE x 2	*14071.2534*	*64150.0906*	*13268.0522*	*14630.2570*	*64734.8578*	*11264.2182*

Tabla 1. Comparativa de resultados Linear Regression vs Random Forest

Como conclusión, si se tuviese que elegir un modelo de los 6 analizados para predecir la potencia, se utilizaría un bosque aleatorio múltiple usando tanto la radiación como la temperatura.

5.1.4 Optimización de hiperparámetros de bosques aleatorios

El modelo de bosque aleatorio dispone de muchos hiperparámetros para personalizar el modelo dependiendo de los datos que se tengan y poder así mejorar el modelo. Este apartado va a consistir en buscar los hiperparámetros óptimos para crear

un mejor modelo, este proceso es conocido como "*fine tuning*". Como se ha mencionado, el modelo sin parámetros se suele adaptar a los datos que se le pasan y se obtienen predicciones muy buenas, pero son mejorables mediante los siguientes hiperparámetros que se van a presentar [34]:

- *n_estimators*: es el número de árboles incluidos en el modelo. Mayor número de arboles aumenta el rendimiento y hace que las predicciones sean más estables. Sin embargo, si se aumenta mucho el número de árboles, la computación es más costosa. El rendimiento del modelo aumenta cuando el número de arboles incrementa y luego se estanca a cierto nivel.
- *criterion*: se trata del criterio utilizado para comparar las estimaciones; puede ser MSE (*Mean Squared Error*) o MAE (*Mean Absolute Error*), por defecto es MSE que el que se debe utilizar para la regresión.
- *max_depth*: es la profundidad máxima que pueden alcanzar los árboles. Este hiperparámetro indica la ruta más larga entre el nodo raíz y el nodo hoja. Por defecto el algoritmo selecciona de manera automática el número de nodos hasta que las hojas tengan menos datos. A medida que aumenta el valor *max_depth*, el rendimiento aumenta inicialmente, pero después comienza a disminuir; esto es debido a que el árbol empieza a sobreajustarse.
- *max_features*: indica el número máximo de características que el bosque aleatorio puede probar en un solo árbol.
- *n_jobs*: son el número de núcleos (*cores*) con los que realizamos el entrenamiento Con este parámetro igual a -1 se utilizan todos los *cores* disponibles.
- *oob_score*: indica si se calcula o no el "*Out of Bag* R^2". Por defecto está en "False" para no incrementar el tiempo de entrenamiento.
- *random_state*: se trata de la semilla (valor entero) para que los resultados sean reproducibles.

Y otros más, que no se van a tratar porque no tienen tanta relevancia como los anteriores. Si que pueden ayudar a mejorar el modelo, pero no son tan significantes

como los anteriormente vistos. Pero es importante saber que estos parámetros existen: *min_sample_split, max_leaf_nodes, min_sample_leaf* y *max_samples*.

En primer lugar, se ha probado a modificar el parámetro del número de árboles (*n_estimators*) del bosque, ya que este parámetro es de los que más importancia suele tener sobre el modelo. Cuantos más árboles disponga el bosque, más robusto será, menos varianza tendrá y mejores predicciones hará el modelo, pero si se aumenta mucho el número de árboles el entrenamiento va a ser mucho más costoso. Para empezar y hacerse una idea de si varía el RMSE si se establece un número concreto de árboles, se han creado los modelos simples y el múltiple con 100 árboles. Y como se puede apreciar en la Tabla 2, los valores de RMSE añadiendo el parámetro "*n_estimators = 100*" son mayores que los valores de RMSE sin parámetros; menos en la radiación.

Pero como se debe utilizar el modelo que menos RMSE tenga, este estudio se va a centrar a partir de ahora en los modelos múltiples, que son los que menos RMSE tienen.

Se ha podido comprobar tanto en la regresión lineal como en la los bosques aleatorios, que el RMSE es ligeramente menor cuando se utiliza radiación y temperatura para predecir, que cuando se utiliza solamente radiación. Este efecto refleja que la temperatura también ayuda a la hora de predecir la potencia, aunque esta variable por sí sola no prediga del todo bien la potencia.

	RF (poa)	*RF (Tcell)*	*RF (poa y Tcell)*
RMSE Sin parámetros	*7315.1285*	*32367.4289*	*5632.1091*
RMSE n_estimators 100	*7310.4826*	*32429.4331*	*5633.8937*
RMSE n_estimators 200	*7314.5333*	*32422.4125*	*5618.4011*

Tabla 2. RMSE vs número de árboles

Se ha comprobado que el RMSE varía con el parámetro "*n_estimators*" añadido y se había comentado anteriormente que cuanto mayor sea el número de árboles, menor debería ser el RMSE. ¿Entonces qué ha pasado? El algoritmo sin parámetros ha utilizado un número de árboles mejor que el añadido (100) para el modelo múltiple. Por lo que se va a intentar volver a modificar de nuevo el número de árboles, estableciéndolo ahora a 200. Como era de esperar, se observa en la Tabla 2 que con 200 árboles se obtienen mejores resultados para el modelo de radiación y temperatura, que cuando no se han añadido parámetros y cuando se utilizaban 100 árboles. Es decir, el RMSE para 200 árboles que es igual a 5618.4011 W es menor que el RMSE con 100 árboles, que es 5633.8937 W.

Aquí se ha podido comprobar la relación inversamente proporcional entren RMSE y número de árboles.

¿Cuántos árboles se deben indicar en el diseño del modelo? Pues se pueden indicar los que se quieran para intentar mejorar el modelo, pero como ya se ha mencionado existe un "*trade off*" entre coste computacional y disminuir el RMSE.

Así que para ver como evoluciona el RMSE cuando se aumenta el número de árboles se va a realizar una función que cree 6 modelos con número de árboles diferentes y que almacene los RMSE de cada modelo. Se ha indicado que el primer modelo disponga de 100 árboles, el segundo 200, y así hasta el último que debe tener 600 árboles.

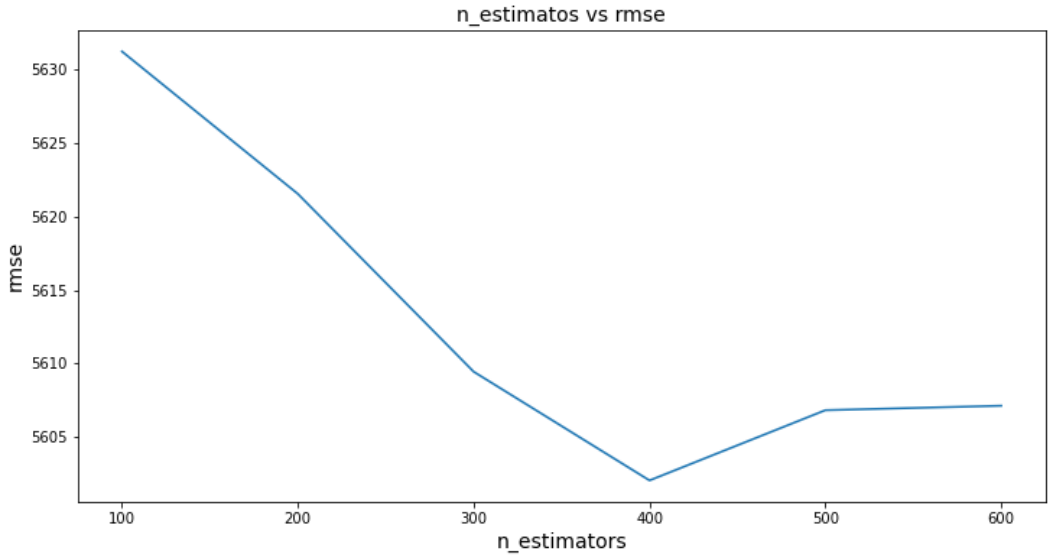

Figura 60. n_estimators vs RMSE

Como se puede observar en la Figura 60, el RMSE disminuye con el aumento de los árboles, aunque llega un punto que no importa cuánto se aumente el número de árboles, que el RMSE se va a mantener. Este efecto se ve claramente en la gráfica ya que entre 100 y 400 árboles el RMSE disminuye a medida que los árboles aumentan, y a partir de los 400 árboles no mejora el modelo, se mantiene el mismo RMSE para 500 y 600 árboles.

Así que, el máximo número de árboles que se debe utilizar es 400, lo cual ya es mucho; ya que en artículos de investigación indican que lo más habitual es utilizar entre 100 y 200 árboles. Entonces, el criterio que se ha seleccionado en este artículo es no seleccionar 400 árboles porque el modelo va a llevar mucho tiempo de entrenamiento, pero tampoco 200 árboles ya que puede ser un poco mejorable. Por lo tanto el máximo de número de árboles que se va a probar en la optimización de parámetros es un valor intermedio, 300 árboles, que permite reducir de manera notable el RMSE.

Una vez se ha estudiado el rango de valores en el que se encuentra el número de árboles, se pasa a buscar una combinación de parámetros óptimos que de cómo resultado un modelo con gran precisión.

Se pueden buscar los mejores hiperparámetros a basa de pruebas manuales, el usuario puede cambiar los parámetros uno a uno y evaluar el modelo, comprobando el

cambio del RMSE. Sin embargo esto es un proceso muy costoso y poco efectivo ya que el mejor modelo puede ser una combinación exacta de varios parámetros simultáneamente. Por eso, se pueden establecer unas listas de valores para cada parámetro y a través de bucles anidados, probar diferentes combinaciones de parámetros. Esto se ha implementado en este estudio incluyendo solo dos parámetros, el número de arboles y el numero máximo de características. Los valores para el número de árboles es desde 5 hasta 200 árboles con un paso de 5 árboles. Es decir: *n_estimators* = 5, 10, 15, ... 190, 195, 200. Y el otro parámetro, solo puede ser 1 o 2 al tener solo 2 variables: radiación y temperatura. Por lo que se ha probado con: *max_features* = 1, 2.

Los parámetros son combinados todos con todos, es decir: el primer modelo tendrá 5 árboles y se probará con *max_features* igual a 1, el siguiente modelo mantendría el número de árboles y probará con *max_features* igual a 2, y así sucesivamente. Cada combinación de 2 parámetros se crea un modelo. Por lo tanto, existen 40 modelos (200/5) para *max_features* igual a 1 y 40 modelos (200/5) para *max_features* igual a 2. Que eso es un total de 80 modelos, cada uno con una combinación de parámetros distinta, por lo que se ha realizado el proceso de entrenamiento y evaluación para cada uno de los 80 modelos. Y se han almacenado las medidas de RMSE y *Out of Bag R²* para cada uno de los modelos.

Al finalizar, se deben buscar los mejores resultados de estas dos medidas para saber qué combinación de parámetros ha sido la óptima. Es decir buscar el valor de RMSE más bajo y el valor de *Out of Bag R²* más alto.

Se ha obtenido que el menor RMSE ha sido: 5535.7017 W. Y la combinación de parámetros que ha dado este resultado es: *n_estimators* = 170 y *max_features* = 1.

Y el valor máximo de *Out of Bag R²* obtenido de entre los 80 modelos es: 0.98883. Para este resultado, la combinación de parámetros ha sido: *n_estimators* = 185 y *max_features* = 1.

Es decir, con los rangos introducidos para esos parámetros, el mejor modelo si se toma como métrica el RMSE, es un bosque aleatorio simple (ya que solo se utiliza una variable), el cual dispone de 175 árboles. Y si se toma como métrica el *Out of Bag R²,* el

mejor modelo es también un bosque aleatorio simple (ya que solo se utiliza una variable), pero este tiene 185 árboles.

Esto ha sido una simple prueba para mostrar que si que se puede diseñar un mejor modelo, ya que el RMSE que se ha conseguido ahora es: 5535.7017 W y el menor RMSE conseguido sin añadir parámetros era: 5632.1091 W. Por lo que si se ha conseguido disminuir el RMSE. Pero esto, ha sido simplemente un experimento para demostrar que si que se puede diseñar una función propia que obtenga parámetros óptimos, pero no es eficaz si se quiere estudiar gran cantidad de parámetros.

Sin embargo existen otras manera de probar los hiperparámetros para un modelo, ya que existen librerías de Python que son más efectivas. Como por ejemplo, la búsqueda de parámetros en cuadricula o también conocido como "*GridSearch*". Otro ejemplo es "*RandomSearch*" o el uso de "*pipelines*" para combinar diferentes modelos.

En este estudio, se van a utilizar conjuntos de hiperparámetros ("*ParameterGrid*") y *GridSearch* para buscar los hiperparámetros óptimos basándose en el "*Out of Bag error*" y en la validación cruzada.

GRIDSEARCH OUT OF BAG ERROR

Out of bag error se trata de una estrategia de validación que permite estimar el error de validación de un modelo sin tener que prescindir de estrategias como la conocida validación cruzada o "*cross validation*", la cual es más costosa computacionalmente hablando.

Como se va a utilizar el mismo "*ParameterGrid*" en las dos métodos de validación, se van a validar los mismos parámetros. Por lo tanto, los parámetros óptimos que se concluyan en una validación deben ser parecidos a la otra validación, pero no tienen por qué ser iguales.

El "*ParameterGrid*" utilizado para la validación *OOB*, va a consistir en 3 parámetros: el número de árboles, el número de características máximo y la profundidad máxima de los árboles. Y los valores que pueden tomar son los siguientes:

'*n_estimators*': [240, 260, 280, 300]

'max_features': [1, 2]

'max_depth': [8, 9, 10, 12, 14]

Este búsqueda de la mejor combinación, consiste en único bucle que recorre todas las combinaciones de parámetros del "ParameterGrid" y entrena el modelo con esos parámetros. Después evalúa los modelos mediante el *Out of Bag R²* y se almacenan estos valores en un *dataframe* para ser ordenados de mayor a menor y así elegir la primera fila, que corresponde con el mayor valor que ha alcanzado el *Out of Bag R².*

Y los resultados obtenidos es que el mayor *Out of Bag R²* es igual a 0.990347 para los siguientes parámetros: *n_estimators*=300, *max_depth*=9, *max_features*=2.

Pero esta prueba se ha realizado varias veces para comprobar si son resultados estables, y se ha comprobado, que también existe otro resultado recurrente que es muy similar. Esto se debe a la aleatoriedad al seleccionar los datos.

El otro resultado obtenido para el *Out of Bag R²* es 0.989902 para los siguientes parámetros: *n_estimators*=300, *max_depth*=8, *max_features*=2. Estos resultados son muy similares ya que solo cambia en una unidad la profundidad máxima de los árboles.

GRIDSEARCH CROSS VALIDATION

A continuación se va a realizar la otra búsqueda de hiperparámetros óptimos a través de la validación cruzada utilizando "*GridsearchCV*".

GridSearchCV debe su nombre a "*Gridsearch cross validation*". Se trata de una función de "sklearn" que realiza una búsqueda completa de los mejores hiperparámetros. A esta función se le pasan los hiperparámetros que resultan interesantes y que pueden mejorar el modelo, y "*GridsearchCV*" se encarga de evaluar todas las combinaciones de parámetros.

Para esta optimización de hiperparámetros se han utilizado los mismos parámetros que optimización anterior, ya que se ha definido el "*ParameterGrid*" igual. Sin embargo esta búsqueda de parámetros se basa en la validación cruzada.

La validación cruzada también conocida como "*cross validation*" es una técnica que consiste en reservar una parte de los datos del conjunto de datos de entrenamiento.

Esa parte reservada no se utiliza para entrenar. Posteriormente se entrena el modelo con los datos de entrenamiento que ya no tienen incluidos los datos de validación. Y por últimos los datos de validación se utilizan para probarlos con el modelo y saber la eficacia que tiene este.

Existen varios métodos de validación cruzada. Uno de ellos es *k-Fold*, el cual sigue un proceso iterativo.

Consiste en dividir los datos de forma aleatoria en k grupos del mismo tamaño, todos los grupos menos uno se utilizan para entrenar el modelo y el conjunto de datos que no se ha utilizado para entrenamiento se utiliza para validación.

Este proceso se repite k veces y en cada iteración se utiliza un grupo distinto como validación. Por lo tanto, al repetirse k veces, se generan k estimaciones del error. Y se promedian estos k errores para tener una estimación del error general.

Se aplica esta técnica de validación cruzada usando como métrica el RMSE, lo cual se debe indicar en el parámetro "*scoring*" de la función "*GridSearchCV*". Así se obtendrá la mejor combinación de parámetros que minimice el RMSE.

Para saber cuáles son los valores de esos hiperparámetros óptimos, se debe consultar utilizando "*best_estimator_*". Al consultarlo, devuelve los siguientes hiperparámetros óptimos: *n_estimators*=300, *max_depth*=8 y *max_features*=2, que corresponden con un RMSE igual a 5252.8788 W.

Y al igual que antes, se ejecuta el código varias veces, y otro de los resultados recurrentes es: *n_estimators*=300, *max_depth*=9 y *max_features*=2, que da como resultado un RMSE igual a 5447.6883 W.

Como se puede observar los parámetros obtenidos para ambas optimizaciones son iguales: *n_estimators*=300, *max_depth*=8 o 9 y *max_features*=2.

Por lo que podemos comparar los dos mejores resultados. En la Tabla 3 se pueden observar los resultados para *n_estimators*=300, *max_depth*=9 y *max_features*=2

	RF (poa y Tcell) *OOB*	*RF (poa y Tcell)* *Validación cruzada*
RMSE	*-*	*5447.6883*
OOB R²	*0.990347*	*-*

Tabla 3. Resultados n_estimators=300, max_depth=9 y max_features=2

Y en la Tabla 4 se pueden observar los resultados para *n_estimators=300, max_depth=8* y *max_features=2*.

	RF (poa y Tcell) *OOB*	*RF (poa y Tcell)* *Validación cruzada*
RMSE	*-*	*5252.8788*
OOB R²	*0.989902*	*-*

Tabla 4. Resultados n_estimators=300, max_depth=8 y max_features=2

Comparando la Tabla 3, con la Tabla 4 se puede observar que la Tabla 3 es la que mejores resultados tiene ya que tiene mayor *OOB R²* y menor RMSE que la Tabla 4. Por lo que finalmente, los hiperparámetros óptimos que se van a utilizar para el modelo son: *n_estimators=300, max_depth=9* y *max_features=2*

5.1.5 Comparativa Regresión lineal y Bosque aleatorio

Se pueden comparar los valores de RMSE para la regresión lineal múltiple sin parámetros y para el bosque aleatorio múltiple, el cual ha sido optimizado con los parámetros: *n_estimators*=300, *max_depth=9* y *max_features*=2. Esta comparativa se

encuentra en la Tabla 5 y se ve que el RMSE para "*Random Forest*" es menor y por lo tanto se trata del mejor modelo de todos para predecir la potencia.

	LR (poa y Tcell)	*RF (poa y Tcell)*
RMSE	*6634.0261*	*5447.6883*

Tabla 5. Resultados LR vs RF

Tras haber estudiado cuál de todos los modelos iba a ser el más preciso para predecir la potencia de una planta solar, se ha llegado a la conclusión que el modelo de bosque aleatorio utilizando las variables de radiación y temperatura y los hiperparámetros mencionados es el mejor para reproducir cómo se comporta la potencia.

5.1.6 Predicción de potencia con los mejores modelos

Por último, se utilizan los mejores modelos para reproducir la potencia que se obtendría un día en concreto. Para ello, se utiliza todos los días anteriores al que se quiere predecir como datos de entrenamiento, y se predice ese día con el modelo ya entrenado. Esta predicción se puede comparar con los datos reales de ese día y ver si la predicción del modelo optimizado es correcta y se asemeja a la real. Si ambas curvas son similares significa que el modelo diseñado es bueno y que los hiperparámetros elegidos son los correctos.

Esta predicción se va a realizar tanto para la regresión lineal múltiple como para el bosque aleatorio múltiple, ya que son los dos mejores modelos. Se ha podido comprobar que para ambos es mejor predecir la potencia con ambas variables.

Para la regresión lineal se va a utilizar los parámetros que vienen por defecto y para el bosque aleatorio se van a utilizar los parámetros obtenidos mediante *GridSearch*.

Esta comparación entre los datos predichos y los valores reales de un día, va a ser representada gráficamente y va a reflejar como de buena sería una predicción de la potencia usando solo los sensores de radiación y temperatura. Esto permite, tener una

idea de la potencia que se puede obtener a largo plazo y por lo tanto realizar cálculos para saber si merece la pena realizar un despliegue para una instalación fotovoltaica.

En la Figura 61 se puede observar la gráfica que compara la potencia real del día 04/05/2013, la cual se representa en color azul y la potencia predicha también para ese día en color naranja. En esta gráfica se puede contemplar cuánto se desvía la realidad de la estimación para un modelo de regresión lineal en el que participan la radiación y la temperatura para predecir la potencia.

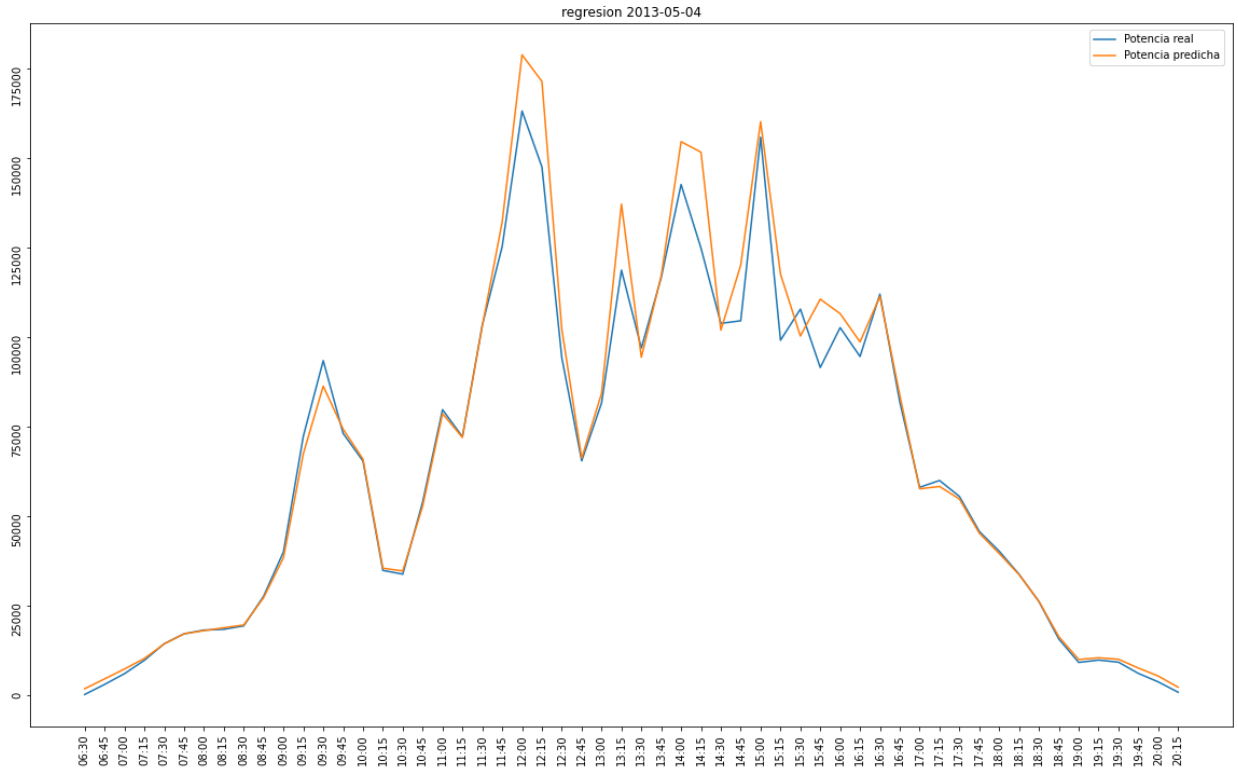

Figura 61. Predecir un día usando regresión lineal

Y la misma gráfica se representa en la Figura 62 para el modelo óptimo de bosque aleatorio. Esta gráfica mantiene la misma correspondencia de colores que la anterior.

Esta gráfica también sirve para visualizar la precisión con la que predice el modelo diseñado. Este modelo se trata de un bosque aleatorio que también predice la potencia utilizando la radiación y la temperatura.

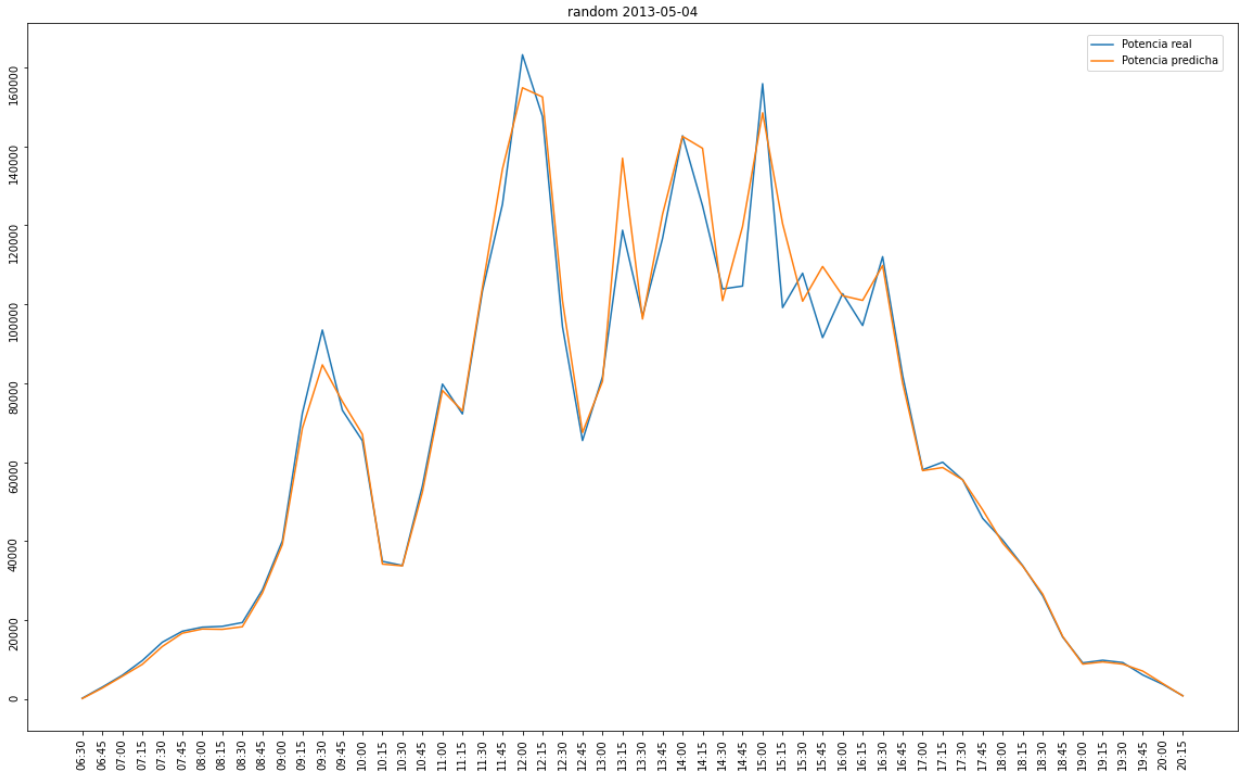

Figura 62. Predecir un día utilizando bosque aleatorio

Como se predice el mismo día en las dos gráficas anteriores para distinto modelo, se puede comparar la Figura 61 con la Figura 62, para ver qué gráfica se desvía más de la potencia muestreada para ese día, y así saber que modelo es mejor para predecir la potencia, si la regresión lineal o el bosque aleatorio.

Y así se ha hecho, llegando a la conclusión que la gráficas son muy parecidas y que los ambos modelos realizan muy buenas predicciones, pero si se tuviese que escoger uno de ellos, sería el modelo de bosque aleatorio en el que la potencia predicha se parece un poco más a la potencia real. Aunque esto ya se podía intuir ya que el RMSE es menor para el modelo bosque aleatorio.

Ya se había mencionado al ver los resultados de la Tabla 5 que el modelo de bosque aleatorio es el que mejor se adapta a los datos de este estudio para predecir la potencia, y esto se reconfirma al observar la comparación entre la Figura 61 y la Figura 62.

Para concluir el apartado de modelización, es interesante destacar el orden de importancia de las variables para predecir la potencia. Este orden es el mismo para la regresión lineal y para el bosque aleatorio.

Ordenados los modelos por orden de más precisos a menos precisos, a la hora de predecir:

1. Radiación y Temperatura (modelo múltiple)
2. Radiación (modelo simple)
3. Temperatura (modelo simple)

5.2 Degradación

La degradación del campo solar se ve reflejada en una disminución del rendimiento del campo solar. Que a su vez, esto implica una disminución de energía producida y por lo tanto de potencia. La degradación se trata de la comparación de rendimiento entre dos instantes diferentes. Y el rendimiento depende de la potencia. Por lo que se está afirmando que la degradación se puede calcular como una comparación o diferencia de potencias.

El problema es que la potencia generada en un campo solar depende sobre todo de 3 factores: radiación, temperatura y tensión. Por lo tanto la potencia, depende del entorno y la climatología. ¿Tiene sentido realizar un experimento en el que se mida la potencia hoy y compararla con la potencia que se mide mañana, y con esa comparación decir si el campo se ha degradado?

No. Lo primero es que la degradación no se puede calcular a corto plazo, se tienen que tener una gran cantidad de datos para obtener unas conclusiones sustanciales.

Lo segundo es que la potencia, como se ha mencionado depende del entorno y esa potencia calculada puede cambiar de un día para otro porque quizás haga más calor, o quizás había nubes, o quizás había más viento y la temperatura del módulo era diferente, o quizás se habían ensuciado los módulos solares, o quizás una combinación de varias de estas situaciones nombradas. Las plantas solares al situarse en el exterior, se

encuentran en entornos que no son estables y la potencia puede ser distinta de un momento a otro sin que se haya producido una degradación.

Entonces, ¿se puede calcular la degradación del campo solar con un solo año?

Se podría pensar en calcular una media de la potencia de cada mes y observar si existe una disminución de potencia entre meses. Ese pensamiento no es correcto, ya que en los meses de invierno no se obtienen las mismas temperaturas que en los meses de verano. Y en los meses de invierno no se obtiene la misma radiación que en los meses de verano ya que el sol "sale" o se "esconde" a diferentes horas. Por lo que las estaciones del año (todas, no solo invierno y verano) influyen en la potencia calculada y no se obtendría un resultado correcto de la degradación.

Puede afectar hasta el nivel de polución a la hora de calcular la potencia, ya que si existen menos capas de la atmosfera o son más débiles, pueden dejar pasar más fácilmente la radiación.

Por lo tanto, la respuesta es que no se puede calcular la degradación con un solo año. Como se ha mencionado, la degradación se calcula a largo plazo. Un año no es suficiente.

De nuevo otra cuestión que se puede plantear, ¿y con varios años, se puede calcular la degradación? Sí. Pero no es simplemente calcular la media de potencia cada año y ver si disminuye para afirmar que existe degradación. Porque los efectos externos y climatológicos que se han explicado anteriormente siguen estando y pueden afectar a la potencia.

Entonces para evitar este problema, se puede buscar en años días que tengan las mismas características de temperatura y radiación y comparar si la potencia disminuye, para confirmar si existe degradación. Pero, ¿y si están actuando otros agentes externos como puede ser el ensuciamiento (*soiling*) del campo solar?

Pues entonces la solución no es seleccionar un día igual en diferentes años con las mismas características climatológicas.

La solución es dividir todos los datos en intervalos de radiación y temperatura con el objetivo de tener todos los datos clasificados en grupos. Y que cada grupo de datos

debe tener valores muy similares de radiación y temperatura (no los mismos valores ya que se han establecido intervalos). Así se podrá estudiar dentro de cada grupo como se comporta la potencia para cada año y ver si esta se degrada. Por tanto, se trata del estudio de la potencia para cada año utilizando datos que tienen características similares de radiación y temperatura.

Ejemplo del estudio de la degradación para un grupo de datos específico (definido en un intervalo) : Estudiar como evoluciona la potencia desde 2013 hasta 2020 utilizando los datos que tengan radiación mayor que 250 $\frac{W}{m^2}$ y menor que 300 $\frac{W}{m^2}$ y una temperatura entre 20ºC y 25ºC.

Este número de intervalos en los que se divide los datos también es conocido como "*bins*". Y el grupo o conjunto de valores que se encuentra dentro del intervalo definido por el *bin* se denomina "*bucket*". Cuanto mayor sea el número de intervalos (mayor *bins*), menor será el número de valores dentro de cada intervalo (menor es el *bucket*). Por lo que los *bins* son inversamente proporcionales a los *buckets*.

Tras esta introducción de los conceptos necesarios para realizar el estudio de la degradación, se va a explicar cómo se desarrolla. El estudio de la degradación se va dividir básicamente entre fases:

1. Filtrar los datos por radiación
2. Dividir los datos en *buckets*
3. Analizar los *buckets*

Es importante que los intervalos estén equitativamente distribuidos, es decir que la cantidad de datos en cada intervalo sea parecida. Esto se debe a que en la fase 3 del estudio de la degradación, se van a realizar cálculos sobre cada *bucket* para analizar la degradación de la potencia. El análisis de un solo *bucket* significa estudiar la degradación de la potencia en un único intervalo de radiación y temperatura determinados.

Como el objetivo del proyecto es obtener la degradación del campo solar para todos los valores de radiación y temperatura a los que está expuesto, se deben tener en cuenta todos los *buckets* que se han definido.

Por lo que, se analizan los valores de todos los *buckets* y se obtienen resultados de cada uno de los *buckets* y luego se promedian los resultados de todos los *buckets*.

Se desea que los intervalos estén equitativamente distribuidos, porque una vez se calcule la potencia de cada *bucket*, se hace una media de los *buckets*. Si un *bucket* tiene pocos valores, los pocos valores que tenga van a influir mucho en la media de todos los *buckets*.

5.2.1 Filtrar los datos por radiación

En esta primera fase del estudio de la degradación de la potencia, se ha decidido seleccionar todos los datos que contengan una radiación mayor que 200 $\frac{W}{m^2}$ (Poa > 200 $\frac{W}{m^2}$) y crear un *dataframe* que contenga solo estos datos, ya que los que no cumplan este requisito no van a ser utilizado en el estudio de la degradación. Esto se ha implementado con una función, por lo que se puede aplicar a los 4 campos solares de forma independiente. Este filtrado de datos se ha realizado para asegurarse de que los datos que se van a utilizar son en condiciones en la que la planta solar obtiene resultados de potencia relevantes, y no se trata de días nublados que pueden aportar ruido al estudio de la degradación. Y este mínimo de radiación elegido 200 $\frac{W}{m^2}$ se ha elegido a partir de los datos del primer cuartil.

Antes de filtrar había 112179 filas y después de haber filtrado por Poa > 200 $\frac{W}{m^2}$, los *dataframes* de los 4 campos solares contienen 72262 filas.

5.2.2 Dividir los datos en buckets

Se consigue que todos los intervalos contengan el mismo número de valores mediante la función "*qcut*". Realmente no tiene que ser exactamente el mismo, sino muy similar. Que los datos estén distribuidos de manera parecida en todos los intervalos se consigue gracias a "*qcut*".

Si se desea tener muchos intervalos (*bins* se define grande), la cantidad de datos en cada *bucket* es pequeño. Entonces, es más probable que existan *buckets* que contengan solo valores que generan ruido, y estos *buckets* van a influir negativamente a la media de todos los *buckets*. Sin embargo, si cada *bucket* contiene mayor número de datos es menos probable que uno de ellos contenga todos sus datos con valores ruidosos.

Sin embargo es conveniente que haya mayor número de *buckets* ya que se consigue clasificar mejor lo datos, es decir en unos intervalos definidos más estrictos donde las características de radiación y temperatura tienen un menor rango. Y los valores de radiación y temperatura se parecen más entre sí dentro de un mismo *bucket*.

Por lo que existe un compromiso a la hora de definir los *bins*, ya que si los *buckets* son muy pequeño, algún *bucket* puede afectar negativamente al estudio de la degradación. Y si los *buckets* son muy grandes no se va conseguir una buena precisión para estimar la degradación.

Así que, en este estudio se ha seleccionado que la radiación se va a dividir en 5 *bins* o intervalos y la temperatura también se va a dividir en 5 *bins* o intervalos.

Primero se va a dividir la radiación en 5 intervalos. Para ello se le pasa a la función "*qcut*" que se necesita dividir los datos filtrados de radiación de la fase 1, en 5 *bins*.

Se puede obtener el limite de los 5 intervalos creados buscando el mínimo y el máximo dentro de cada intervalo. Además también se puede obtener el número de valores de cada intervalo utilizando "*value_counts*".

Para el caso de estudio en cuestión, los resultados que se han obtenido se pueden visualizar en la Figura 63. Se puede apreciar que encima de cada *bucket* hay un valor a la izquierda y otro a la derecha, siendo estos los valores límite.

El primer *bucket* (*bucket* 0), este contiene valores que van desde 200.01 $\frac{W}{m^2}$ hasta 343.25 $\frac{W}{m^2}$.

También se puede observar en esta Figura 63, que se muestran la cantidad de valores que contiene cada intervalo, y son muy similares entre ellas; así como se requería.

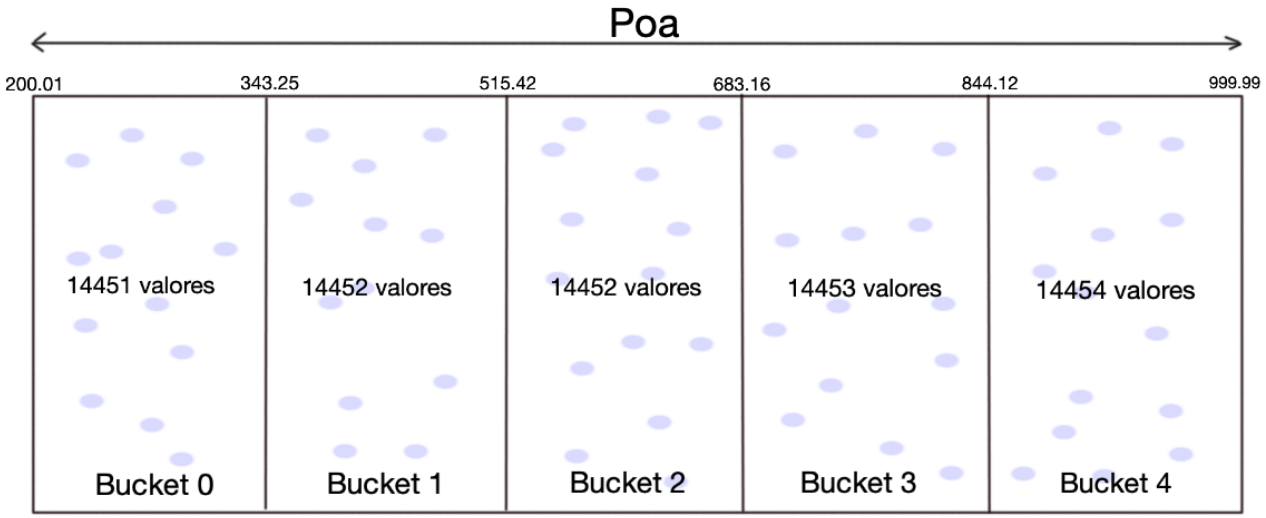

Figura 63. Buckets poa

Si se hace la suma de la cantidad de valores de cada *bucket* debería dar igual que el número de filas tras haber realizado el filtrado de la radiación: 72262 filas. Y así se cumple.

A continuación se va a realizar la división de los intervalos de la temperatura. Los *bins* de Tcell también se ha definido como 5. Así que cada *bucket* de radiación ya creado se va a dividir en 5 intervalos de temperatura. Dando lugar a 25 *buckets* como se indica en la Figura 64.

Para implementar esta estructura de intervalos en Python, se debe seleccionar los datos contenidos en intervalo de radiación y estos datos se pasan a la función "*qcut*" para que los divida en 5 intervalos de temperatura.

Y este proceso se debe realizar 5 veces, una por cada *bucket* de "poa".

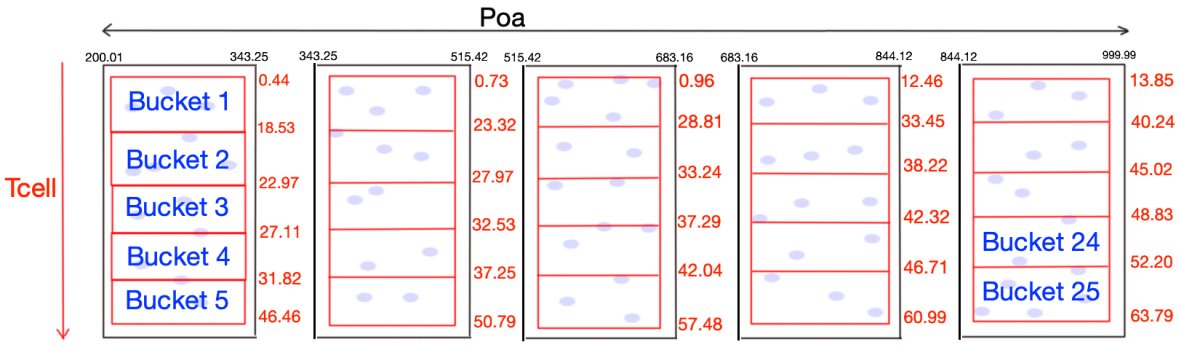

Figura 64. Buckets Tcell en cada bucket poa

Como los datos de cada *bucket* de radiación son distintos, cuando estos datos se clasifican según temperatura, se van a obtener diferentes limites en los intervalos. Es decir, el *bucket* 0 de radiación no tiene por qué tener los mismos limites en los intervalos de temperatura que el *bucket* 1. Y así con los 5 *buckets*.

Esto se puede comprobar en la Figura 65, donde se puede observar que el primer *bucket* de temperatura en el primer *bucket* de radiación tiene los siguientes límites: 0.44ºC - 18.53ºC. Y en cambio el primer *bucket* de temperatura en el segundo *bucket* de radiación tiene otros límites: 0.73ºC - 23.32ºC. Esto es un ejemplo, pero esto ocurre en todos los *buckets*.

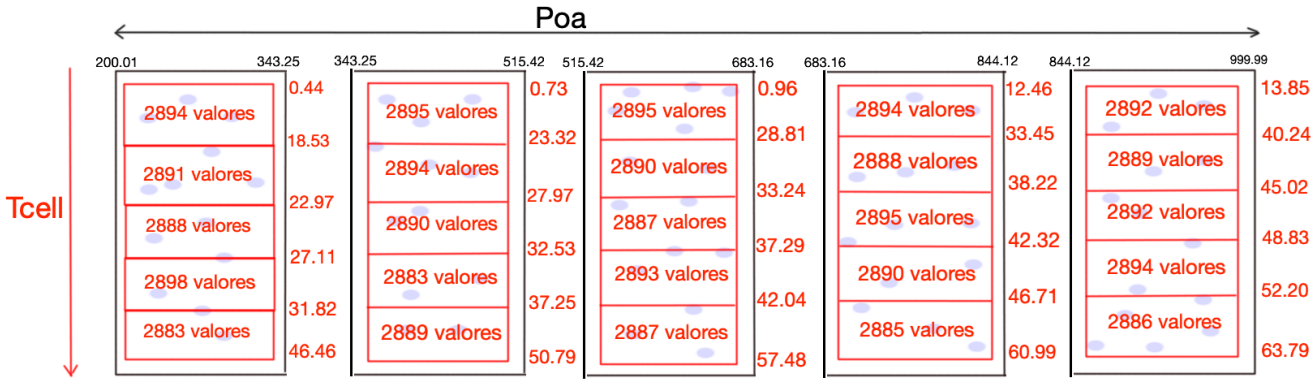

Figura 65. Cantidad de valores en los 25 buckets

Esta distribución queda como una tabla, en la que los intervalos de radiación podrían ser las columnas y los intervalos de temperatura las filas. Pero como los intervalos

de temperatura no son fijos, ya que dependen del intervalo de radiación en el que estén; realmente sería como se tuviesen 5 tablas. Donde cada celda de la tabla corresponde con un intervalo.

También hay que tener en cuenta que la Figura 65 no está dibujada teniendo en cuenta los límites de los intervalos de temperatura. Ya que no todos los intervalos de temperatura tienen el mismo rango, y sin embargo se han representado como si todos fueran iguales.

Además, en la Figura 65 también se presenta el número de valores que tiene cada uno de los 25 *buckets*. Se puede comprobar que todos son muy parecidos. Y también se puede comprobar que si se suman lo cantidad de valores de los 25 *buckets* también debe dar igual que el número de filas tras haber realizado el filtrado de la radiación: 72262 filas.

Ahora que ya se han distribuido los datos en grupos o *buckets*, donde los datos se clasifican por similar radiación y temperatura, se continua con el análisis de cada *bucket* por separado y posteriormente del conjunto de *buckets*.

5.2.3 Analizar los buckets

A modo de resumen, en esta fase se va a realizar lo siguiente:

Para cada *bucket* generado en la fase anterior (pareja de *bucket* poa y *bucket* tcell) calcular la mediana de la potencia año por año (2013-2020). Y se va a hacer una gráfica para visualizar la evolución de la potencia. En esta misma fase, también se va calcular la diferencia de potencia entre un año y el siguiente. Y además se va a comprobar si la potencia entre el primer año y el último excede el 1%.

Se pasa a ahora a explicar en detalle la tercera fase que consiste en analizar los *buckets* generados:

Una vez que todos los intervalos contienen un número de valores similares, se calcula la mediana de la potencia para todos los años dentro de un mismo intervalo. Esto significa, calcular la mediana de la potencia para todos los años dentro de un rango o intervalo de radiación y temperatura similares.

Para seleccionar los valores de cada intervalo, se debe utilizar dos bucles anidados, el primer bucle para acceder a los intervalos de radiación y el siguiente bucle para acceder a los intervalos de temperatura dentro de cada intervalo de radiación.

Cuando se tienen seleccionados todos los valores de un *bucket* se dividen los datos por años, desde 2013 hasta 2020, obteniendo así 8 grupos de datos en 1 *bucket* porque hay 8 años. Y finalmente se calcula la mediana de la potencia para cada uno de los 8 grupos de datos, es decir la mediana de la potencia para cada año. Obteniendo así 8 valores "*float*", que representa la mediana de la potencia de cada año en ese intervalo de radiación y temperatura específico.

Como se puede observar en la Figura 66, del primer *bucket* se obtienen 8 valores. Cada uno de ellos es la mediana de la potencia en un año, y está indicado como "pow + año" en color verde. Este análisis de años y de potencia se realiza en los 25 *buckets*, pero en la Figura 66 solo se ha representado el primer *bucket* y el último para entender la idea. El resto de *buckets* están representados como puntos suspensivos.

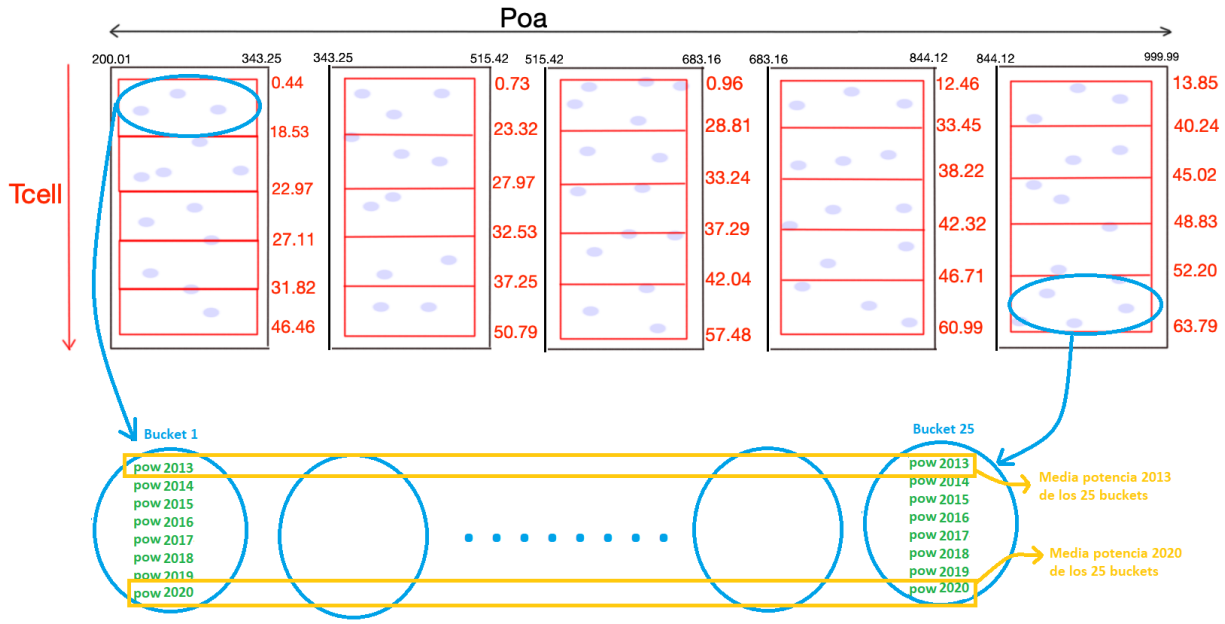

Figura 66. Análisis de buckets: obtener la potencia

Una vez se obtiene la mediana de la potencia para cada año en cada intervalo, se realiza la media de los resultados de los 25 intervalos. Es decir, se van a seleccionar los resultados de potencia de 2013 de todos los intervalos y se va a calcular la media de estas 25 potencias. Luego, se van a seleccionar los resultados de potencia de 2014 de todos los intervalos y se va a calcular la media de estas 25 potencias. Y así sucesivamente.

Esto proceso se puede observar en la Figura 66, donde se seleccionan los 25 valores con un rectángulo en amarillo y se obtiene la media de todas esas potencias para ese año.

Se ha representado este proceso de obtener la media de todos los *buckets*, solo para el primer año (2013) y para el último año (2020); aunque realmente se tiene que hacer para todos los años. Es decir se obtienen 8 medias, una por cada año.

Y ahora que se tiene la potencia media de cada año, se puede representar en la Figura 67 como evoluciona la potencia.

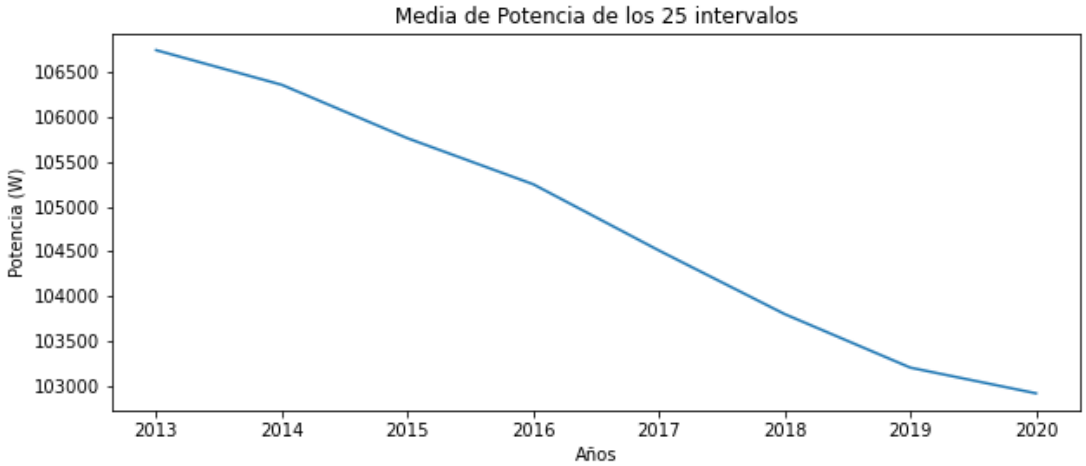

Figura 67. Evolución de la potencia por años

Se puede observar cómo esta potencia va disminuyendo con el paso de los años, y este efecto no es debido a la baja radiación debido a las nubes (la radiación se ha filtrado), ni se debe a cambios de la radiación o temperatura ya que estas han sido clasificadas en intervalos (para eso se han creado los *buckets*). Por lo que en la Figura

67, se puede apreciar claramente que existe degradación del campo solar con el paso de los años.

Sería interesante saber cuanto se ha degradado el campo solar desde el 2013 hasta el 2020. Así que se va a calcular la diferencia de potencias entre cada año y también entre el primer año y el último, para saber la degradación total en todos estos años.

La diferencia que se calcula entre años, se podría ver muy influenciada si por ejemplo un año ha habido más nubes que otro. Por eso se utilizan los *buckets*. Así, para cada *bucket* se obtiene una diferencia porcentual de medianas.

El proceso para obtener las diferencias de potencias entre años es muy parecido al anterior de la potencia. En cada *bucket*, se dividen los datos por años como se ha explicado y posteriormente se calcula la mediana para cada año.

Si se representan estas medianas de potencia para el primer *bucket* se obtiene la Figura 68. Y se puede calcular la diferencia entre dos años (marcado en verde en la Figura 68) siguiendo la siguiente formula:

$$\text{diferencia}=\text{abs}\left(\frac{\text{potencia año actual}-\text{potencia año siguiente}}{\text{potencia año actual}}\right)\text{x}100\ \ (\%)$$

Se trata de la diferencia del año actual respecto al siguiente. Como esta puede ser negativa se obtiene el valor absoluto y por último se multiplica por 100 para obtenerla como una diferencia porcentual.

Y para la diferencia entre el primer y el ultimo año sería:

$$\text{diferencia}=\text{abs}\left(\frac{\text{potencia año actual}-\text{potencia último año}}{\text{potencia año actual}}\right)\text{x}\frac{1}{7}\text{x}100\ \ (\%)$$

Se divide entre 7 porque hay 7 diferencias, y lo que se quiere obtener es la degradación anual en porcentaje entre el primer año y el último.

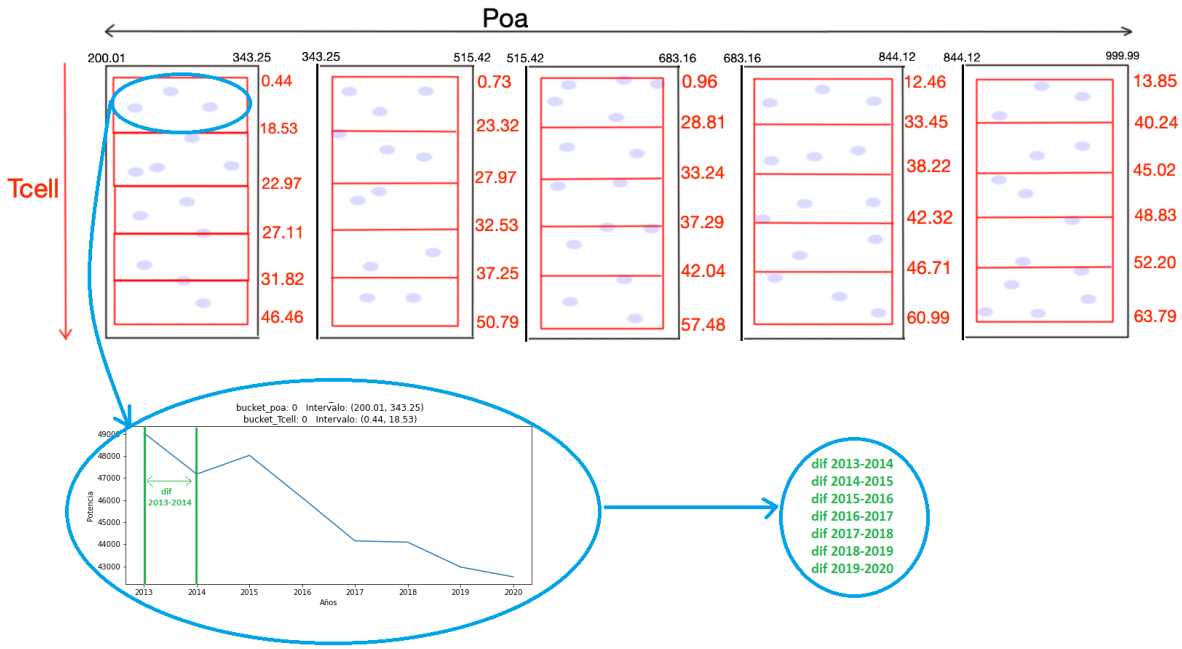

Figura 68. Diferencia de potencia por años

Antes, cuando se quería obtener la potencia se obtenían 8 resultados de cada *bucket*, porque había 8 años. Ahora como se está obteniendo la diferencia entre años, solo se van a obtener 7 resultados, uno por cada diferencia: dif 2013-2014, dif 2014-2015, dif 2015-2016, dif 2016-2017, dif 2017-2018, dif 2018-2019, dif 2019-2020. Las cuales se pueden observar en color verde en la Figura 68, haciendo referencia a los 7 niveles de degradación medidos en porcentaje.

Estas 7 diferencias son calculadas para los 25 *buckets*. Aunque en la Figura 69 solo se represente para el primer y último *bucket*. Y posteriormente se hace la media de los 25 *buckets* para cada diferencia de años.

Como resultado final se debe obtener 7 medias. Y cada media corresponde a una diferencia entre dos años.

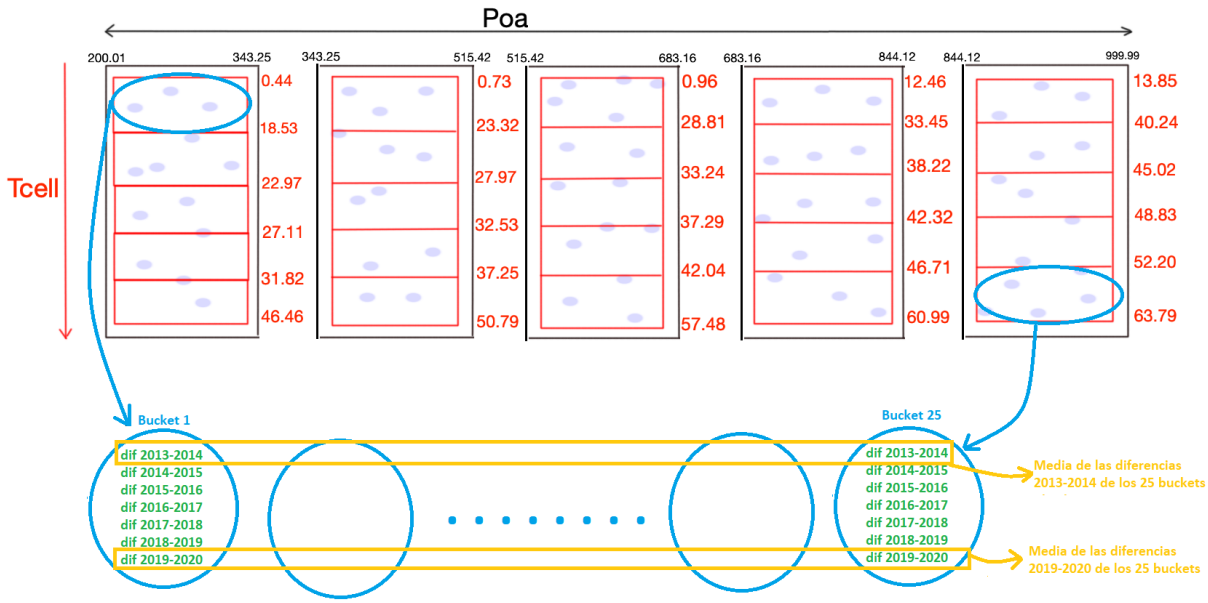

Figura 69. Análisis de buckets: obtener las diferencias

Estas 7 diferencias de potencia entre años son las siguientes:

Diferencia potencia entre 2013-2014: 1.61682%

Diferencia potencia entre 2014-2015: 1.60824%

Diferencia potencia entre 2015-2016: 1.54072%

Diferencia potencia entre 2016-2017: 1.77491%

Diferencia potencia entre 2017-2018: 1.75974%

Diferencia potencia entre 2018-2019: 1.82062%

Diferencia potencia entre 2019-2020: 1.94034%

Y estas diferencias de potencia han sido representadas gráficamente en la Figura 70.

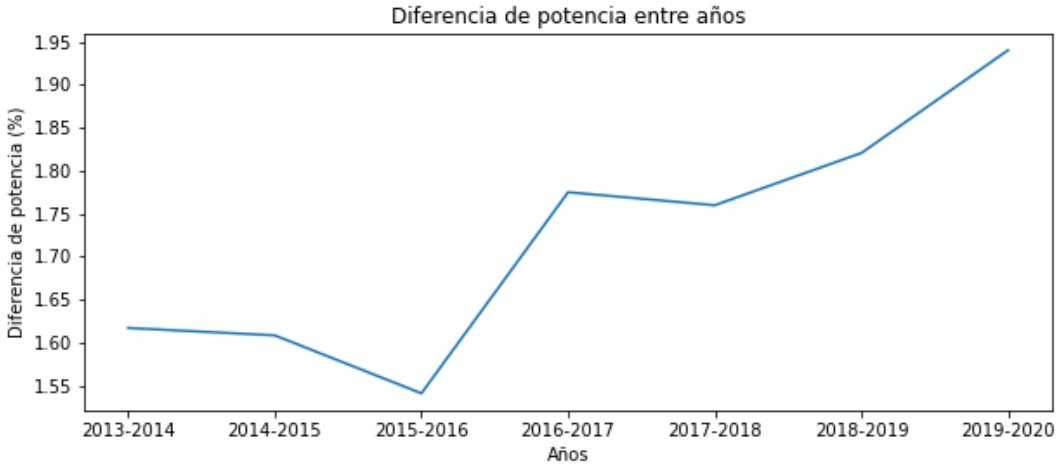

Figura 70. Diferencias de potencia

La diferencia de potencia entre dos años, cada vez debe ser mayor a medida que avanzan los años, ya que la degradación del campo solar no es lineal, sino que con el paso del tiempo, cada vez se degrada más.

Se puede ver claramente como aumenta la diferencia de potencia a lo largo de los años, lo que significa que el campo solar se va degradando cada vez más.

Aunque es cierto que la diferencia de potencia entre 2015 y 2016 es pequeña, dando a entender que entre estos dos años no se observa degradación.

Además, como se había mencionado anteriormente, también se ha calculado el porcentaje de degradación entre el primer año, que es 2013 y el último, que es 2020. Se sigue el mismo procedimiento que para los dos casos anteriores (potencias y diferencias): se obtiene la diferencia entre 2013 y 2020 para cada *bucket* y posteriormente se hace la media de los resultados de los 25 *buckets*, y se obtiene la degradación final.

Las diferencias de potencia porcentuales entre 2013 y 2020 para cada *bucket* son las siguientes:

1.9020351668670599, 1.9463780138307505, 1.419471732648334, 0.4316276733339879, 0.9034884419119471, 1.0572955828316606, 0.7982122664138795, 0.13464080249513746, 0.3545127236036318, 0.8986669490567885, 0.8586502369193714, 0.9913314392532656, 0.09839352259867495, 0.08309653824169776, 1.0722218222049427, 0.789296994182414,

1.190173887641649, 0.4672234326863726, 0.5710216659686789, 1.3302909650457397, 0.453582037444077, 0.7667990064249366, 0.585245409811393, 0.5886515940507104, 1.264285002016101

Y haciendo la media de todas estas diferencias, se obtiene que la degradación entre el primer y último año es : **0.83826%** al año.

Esto no significa que desde el año 2013 hasta el año 2020 se haya degradado el campo solar un 0.83826%, sino que la media de la degradación entre el año 2013 y 2020 es un 0.83826% al año.

En efecto, esta degradación es menor que el 1%, como era de esperar. Ya que los fabricantes de paneles suelen garantizar que en los 10 primeros años deben proporcionar al menos el 90% de la potencia nominal. Si los paneles se degradasen más de un 1% anualmente, a los 10 años se habrían degradado más de un 10%. Y por lo tanto los fabricantes no cumplirían la garantía establecida.

Por eso, la degradación de los campos solares entre dos años suele ser de aproximadamente un 0.5%, como ya se había mencionado anteriormente. Por lo que el resultado obtenido se trata de una degradación un poco superior a lo habitual, pero cumple la garantía.

Aunque los campos solares no den los mismos resultados de potencia, la degradación ha sido en los 4 casos muy similar:

Campo solar 2: 0.66570%

Campo solar 3: 0.60338%

Campo solar 4: 0.88706%

La empresa ha garantizado que el mantenimiento de su planta solar ha sido el correcto, limpiando los paneles solares con bastante frecuencia para que no afecte el efecto de ensuciamiento en el rendimiento de los paneles. Por lo que en este estudio, el ensuciamiento de los paneles no va a afectar a la potencia generada y por lo tanto no se ha de tener en cuenta en el estudio de la degradación.

Capítulo 6 - Conclusiones y trabajo futuro

Como se ha podido comprobar en el capítulo 5.1.5 se ha conseguido crear un modelo muy preciso utilizando un bosque aleatorio múltiple con parámetros optimizados que tan solo presenta un RMSE de 5447.6883 W, los que indica que se puede predecir la potencia a través de la radiación y la temperatura con una gran precisión. Esto se puede apreciar en el capítulo 5.1.6 cuando se representa gráficamente una comparativa de la potencia real con la potencia predicha, ya que se observa que la potencia predicha se superpone a la potencia real en casi toda su evolución.

Aunque se trata de un resultado particular para una planta solar concreta, el resultado nos hace ser optimistas con respecto al primer objetivo del trabajo: predecir la potencia usando sensores de radiación y temperatura usando modelos de aprendizaje automático.

Si este resultado pudiera confirmarse con datos de otras plantas, lo que se deja como trabajo futuro, se tendría que mediante sensores se puede predecir con gran precisión la energía que puede generar la planta en el lugar indicado, lo que resulta de gran utilidad.

En cuanto al segundo objetivo del trabajo: estudiar la degradación de los campos solares de una empresa desde 2013 hasta 2020, hemos planteado un método basado en el estudio de intervalos de características similares (radiación y temperatura), lo que ha permitido estudiar la degradación de la potencia con el paso de los años minimizando la influencia de la climatología.

En particular, en el capítulo 5.2, se calcula la diferencia de la potencia entre años y se puede ver como la degradación va aumentando con el paso de los años. Y mediante otra gráfica, se llega también a la misma conclusión ya que la potencia va disminuyendo notablemente con el tiempo.

Además de comprobar la degradación que existe entre dos años consecutivos, también se ha calculado la degradación entre el primer (2013) y el último año (2020), y se obtiene que el campo solar se ha degradado un **0.83826%** al año desde el primer al último año.

Se puede concluir afirmando que se han cumplido los dos objetivos de este Trabajo Final de Máster, ya que se ha conseguido realizar un estudio de la degradación del campo solar y una predicción de la potencia o rendimiento que tendría una planta solar a través de la radiación y la temperatura.

Chapter - Introduction

This project will explain how solar energy works at large ranges, providing theoretical concepts that are necessary for the study of the degradation of the solar field. In addition, we will explain how the practical study has been implemented in Python, explaining in more detail the process followed in each phase. This experiment will consist of using models to predict the performance of solar fields and using massive data processing techniques to obtain results of the magnitude with which the solar field degrades.

MOTIVATION

In recent years, renewable energies are continuously growing, and this is expected to continue. The impact of non-renewable energies on the environment is causing them to be used less and less.

Among the renewable energies, one of the most outstanding is solar energy, which is increasing its participation in the generation of electrical energy.

The material needed to create solar energy is becoming more and more accessible, encouraging people to install solar modules in their homes. In addition, in recent years the price of solar modules has been decreasing to promote this energy.

As the modules degrade over the years, it is advisable to select modules of good materials at the time of purchase, otherwise the modules may degrade faster.

The manufacturers of solar modules must comply with a performance guarantee where it indicates that the module will generate minimum performance values.

In order to pollute as little as possible, modules should be purchased with the longest lifetime and the least degradation in order to generate as little waste as possible.

To verify that the manufacturer's warranty is met, it is necessary to perform a study of the degradation of the solar field.

Although these studies are not usually carried out in domestic solar plants, they are usually carried out in large companies that collaborate with the country's electricity grid.

These companies usually have large solar plants that generate a lot of power and it is interesting for them to know the degradation of the solar field. Since the loss of performance in large solar plants causes a very significant economic loss.

Therefore, solar power generating companies study the degradation of the purchased modules to know if the manufacturer complies with the warranty and if it does not comply claim to replace them.

So these companies have monitoring systems to sample data from the solar plant and thus be able to perform studies of degradation or predictions of plant performance in the future. Which is very interesting for these companies.

And in this Master's Final Project we are going to perform this study of the degradation of the solar field and a predictive study of its performance with the data of a company.

OBJECTIVES

The objective of this Master's thesis is to perform a study of the degradation of the solar field using the large amount of data provided by a company. To carry out this practical study, power will be used to understand how the performance of the solar field evolves over the years and thus evaluate the degradation between two years or the degradation between the first and the last year. This study will require the use of data processing and analysis techniques to draw conclusions from the more than five million pieces of data that the company has provided for this project. But to understand this study of degradation it is necessary to present some previous theoretical knowledge about how photovoltaic systems work.

In short, the Master's Final Project can be broken down into two specific objectives, which can be summarized as:

-Predict power using radiation and temperature sensors using machine learning models.

-Study the degradation of a company's solar fields from 2013 to 2020.

STRUCTURE

This paper is structured in 5 chapters in total . The first one consists of a preamble of the present Master Final Paper, and then chapter two is presented, which is a general introduction of solar energy . After this theoretical introduction, chapter three shows the most important theoretical concepts that are necessary to analyze the degradation of the solar field.

Chapter four presents the practical study implemented with emphasis on the 4 phases that have been developed: preprocessing, descriptive statistics, machine learning and degradation study. Finally, the fifth chapter presents the conclusions obtained.

Chapter - Conclusions and future work

As can be seen in chapter 5.1.5, a very accurate model has been created using a multiple random forest with optimized parameters that only has an RMSE of 5447.6883 W, which indicates that the power can be predicted through radiation and temperature with high accuracy. This can be appreciated in chapter 5.1.6 when a comparison of the real power with the predicted power is graphically represented, since it is observed that the predicted power overlaps with the real power in almost all its evolution.

Although this is a particular result for a particular solar plant, the result makes us optimistic about the first objective of the work: to predict the power using radiation and temperature sensors using machine learning models.

If this result could be confirmed with data from other plants, which is left as future work, it would be possible to predict with great accuracy the power that can be generated by the plant in the indicated place, which is very useful.

Regarding the second objective of the work: to study the degradation of the solar fields of a company from 2013 to 2020, we have proposed a method based on the study of intervals of similar characteristics (radiation and temperature), which has allowed us to study the degradation of the power over the years minimizing the influence of the weather.

In particular, in chapter 5.2, the difference in power between years is calculated and it can be seen how the degradation increases over the years. And through another graph, the same conclusion is also reached as the power is decreasing significantly over time.

In addition to checking the degradation between two consecutive years, the degradation between the first (2013) and the last year (2020) has also been calculated, and it is obtained that the solar field has degraded **0.83826%/year** from the first to the last year.

It can be concluded by stating that the two objectives of this master's final project have been fulfilled, since it has been possible to carry out a study of the degradation of

the solar field and a prediction of the power or performance that a solar plant would have through radiation and temperature

BIBLIOGRAFÍA

[1] «SOLAR - ENERGÍAS RENOVABLES E INGENIERÍA», *grupoibal*. http://grupoibal.com/solar/ (accedido sep. 04, 2021).

[2] «Infographic: Where Solar & Wind Power Are Thriving», *Statista Infographics*. https://www.statista.com/chart/22558/wind-and-solar-generation-in-selected-countries/ (accedido sep. 04, 2021).

[3] T. text provides general information S. assumes no liability for the information given being complete or correct D. to varying update cycles y S. C. D. M. up-to-D. D. T. R. in the Text, «Topic: Solar energy in Spain», *Statista*. https://www.statista.com/topics/8058/solar-energy-in-spain/ (accedido ago. 30, 2021).

[4] E. texto proporciona información general S. no se hace responsable de la veracidad o exactitud del contenido N. ciclos de actualización varían y D. M. Q. L. E. P. C. I. M. A. Q. L. R. E. E. Texto, «Tema: Las energías renovables en España», *Statista*. https://es.statista.com/temas/6675/las-energias-renovables-en-espana/ (accedido ago. 30, 2021).

[5] «Qué haremos con todos los paneles solares cuando terminen su vida útil». https://ovacen.com/paneles-solares-vida-util/ (accedido ago. 30, 2021).

[6] «Los mayores fabricantes de placas solares», *Otovo Blog*, ago. 29, 2021. https://www.otovo.es/blog/placas-solares/fabricantes-de-placas-solares/ (accedido sep. 01, 2021).

[7] Eduardo Lorenzo, *Radiación solar y dispositivos fotovoltaicos*, vol. 2. 2006.

[8] WPadmin, «SENSORES DE TEMPERATURA AMBIENTAL | Solar Irradiance Sensor – PV Reference Cell – Solar Measurement Equipment – IMT Solar». https://imtsolar.com/sensores-de-temperatura/ (accedido sep. 01, 2021).

[9] «Célula Fotovoltaica: Lingotes y Obleas de silicio solar.» https://www.sfe-solar.com/noticias/articulos/celula-fotovoltaica-lingotes-obleas/ (accedido ago. 30, 2021).

[10] «Diferencias entre silicio monocristalino y multicristalino o policristalino». https://autosolar.es/blog/placas-fotovoltaicas/diferencias-entre-silicio-monocristalino-y-multicristalino-o-policristalino (accedido ago. 30, 2021).

[11] «Fig. 2. – The value chain for the fabrication of monocrystalline (top)...», *ResearchGate*. https://www.researchgate.net/figure/The-value-chain-for-the-fabrication-of-monocrystalline-top-and-multicrystalline_fig2_318665333 (accedido sep. 04, 2021).

[12] «FABRICACIÓN DE PANELES FOTOVOLTAICOS, EFECTOS CONTAMINTES Y RECICLADO.», p. 29.

[13] M. H. Rashid, *Power Electronics Handbook*. Butterworth-Heinemann, 2017.

[14] «(PDF) An Overview of Factors Affecting the Performance of Solar PV Systems». https://www.researchgate.net/publication/319165448_An_Overview_of_Factors_Affecting_the_Performance_of_Solar_PV_Systems (accedido ago. 30, 2021).

[15] «Figure 2. The power versus voltage curves: (a) for different solar...», *ResearchGate*. https://www.researchgate.net/figure/The-power-versus-voltage-curves-a-for-different-solar-irradiation-b-for-different_fig2_352078653 (accedido sep. 04, 2021).

[16] «Orientación e inclinación de placas solares para un máximo rendimiento», *Otovo Blog*, jun. 04, 2021. https://www.otovo.es/blog/placas-solares/orientacion-e-inclinacion-placas-solares/ (accedido ago. 30, 2021).

[17] M. R. Maghami, H. Hizam, C. Gomes, M. A. Radzi, M. I. Rezadad, y S. Hajighorbani, «Power loss due to soiling on solar panel: A review», *Renew. Sustain. Energy Rev.*, vol. 59, pp. 1307-1316, jun. 2016, doi: 10.1016/j.rser.2016.01.044.

[18] Å. Skomedal, H. Haug, y E. S. Marstein, «Endogenous Soiling Rate Determination and Detection of Cleaning Events in Utility-Scale PV Plants», *IEEE J. Photovolt.*, vol. 9, n.º 3, pp. 858-863, may 2019, doi: 10.1109/JPHOTOV.2019.2899741.

[19] R. R. Cordero *et al.*, «Effects of soiling on photovoltaic (PV) modules in the Atacama Desert», *Sci. Rep.*, vol. 8, n.º 1, p. 13943, sep. 2018, doi: 10.1038/s41598-018-32291-8.

[20] D. C. Jordan y S. R. Kurtz, «Photovoltaic Degradation Rates-an Analytical Review: Photovoltaic degradation rates», *Prog. Photovolt. Res. Appl.*, vol. 21, n.º 1, pp. 12-29, ene. 2013, doi: 10.1002/pip.1182.

[21] «Enemigos de la Fotovoltaica - Efecto LID», *Amara-e*, abr. 29, 2020. https://www.amara-e.com/efecto-lid-fotovoltaica/ (accedido ago. 30, 2021).

[22] T. Sun, «El efecto PID (Potential Induced Degradation) en paneles solares», *Techno Sun - Distribuidor mayorista*, sep. 24, 2020. https://www.technosun.com/es/blog/efecto-pid-paneles-solares/ (accedido sep. 06, 2021).

[23] «Solar PV cell construction», *CLEAN ENERGY REVIEWS*. https://www.cleanenergyreviews.info/blog/solar-pv-cell-construction (accedido ago. 30, 2021).

[24] G. Meyer, «El presente y futuro deparan a la tecnología fotovoltaica grandes retos en cuanto a la calidad de material, instalación, operación y, finalmente, desmantelamiento. La fotovoltaica es una tecnología madura, pero solo si el sector responde adecuadamente a las exigencias de suministro seguro y fiable, puede establecerse como fuente principal de electricidad a gran escala.», p. 2.

[25] C.-12 Foundation, «Diagramas de Caja y Bigotes | CK-12 Foundation». https://flexbooks.ck12.org/cbook/ck-12-%c3%81lgebra-i-en-espa%c3%b1ol/section/11.8/primary/section/diagramas-de-caja-y-bigotes (accedido sep. 04, 2021).

[26] G. M. K, «Machine Learning Basics: Decision Tree Regression», *Medium*, jul. 18, 2020. https://towardsdatascience.com/machine-learning-basics-decision-tree-regression-1d73ea003fda (accedido ago. 30, 2021).

[27] G. M. K, «Machine Learning Basics: Decision Tree Regression», *Medium*, jul. 18, 2020. https://towardsdatascience.com/machine-learning-basics-decision-tree-regression-1d73ea003fda (accedido sep. 04, 2021).

[28] «(4) APRENDIZAJE SUPERVISADO: DECISION TREE REGRESSION | #8 Curso de Introducción a Machine Learning - YouTube».

https://www.youtube.com/watch?v=tnDVtjtYDSY&list=LL&index=37 (accedido sep. 04, 2021).

[29] «Decision Tree Regressor explained in depth», *GDCoder*, may 23, 2019. https://gdcoder.com/decision-tree-regressor-explained-in-depth/ (accedido sep. 04, 2021).

[30] S. Kumar, «3 Techniques to avoid Overfitting of Decision Trees», *Medium*, jun. 01, 2021. https://towardsdatascience.com/3-techniques-to-avoid-overfitting-of-decision-trees-1e7d3d985a09 (accedido ago. 30, 2021).

[31] «Decision Tree Ensembles- Bagging and Boosting | by Anuja Nagpal | Towards Data Science». https://towardsdatascience.com/decision-tree-ensembles-bagging-and-boosting-266a8ba60fd9 (accedido ago. 30, 2021).

[32] J. O. A.- johanna.orellana@ucuenca.edu.ec, *Arboles de decision y Random Forest*. Accedido: sep. 04, 2021. [En línea]. Disponible en: https://bookdown.org/content/2031/ensambladores-random-forest-parte-i.html

[33] «randomforest2001.pdf». Accedido: ago. 30, 2021. [En línea]. Disponible en: https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf

[34] O.-O. D. Science, «Optimizing Hyperparameters for Random Forest Algorithms in scikit-learn», *Medium*, jul. 26, 2019. https://medium.com/@ODSC/optimizing-hyperparameters-for-random-forest-algorithms-in-scikit-learn-d60b7aa07ead (accedido ago. 30, 2021).